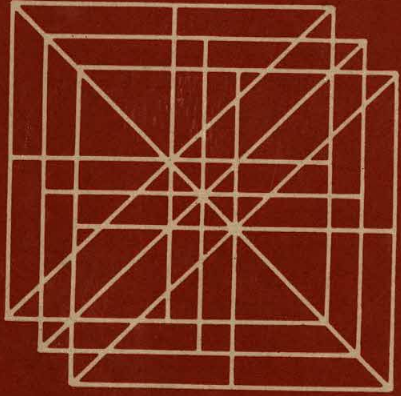
INTRODUCTION TO MODERN ALGEBRA

FOURTH EDITION



NEAL H. McCOY
REVISED BY
GERALD J. JANUSZ

✓

১৮৪৭

UBSPD

UBS Publishers' Distributors Ltd
8/1-B, Chowringhee Lane, Calcutta-700 016
Ph:212451, 2 94/3

Introduction to Modern Algebra

Introduction to Modern Algebra

Fourth Edition

NEAL H. McCOY
Professor Emeritus of Mathematics
Smith College

Revised by

GERALD J. JANUSZ
Professor of Mathematics
University of Illinois

UNIVERSAL BOOK STALL
NEW DELHI

Copyright © 1989 by Wm. C. Brown Publishers. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

First Indian Reprint 1990

For sale in India only.

Universal Book Stall
5 Ansari Road, New Delhi-110 002

ISBN 81-85392-71-4

Acc. no - 16691

Printed at Pearl Offset Press, 5/33, Kirti Nagar Industrial Area, New Delhi-110 015

Dedicated to the memory of my son

PAUL

Neal H. McCoy

Contents

Preface xi

Introduction: The Logical Structure of Mathematics 1

Chapter 1. Some Fundamental Concepts 6
- 1.1 Sets 6
- 1.2 Mappings 12
- 1.3 Equivalence Relations 17
- 1.4 Operations 20

Chapter 2. Rings 23
- 2.1 Formal Properties of the Integers 23
- 2.2 Definition of a Ring 24
- 2.3 Examples of Rings 26
- 2.4 Some Properties of Addition and Multiplication 34
- 2.5 General Sums and Products 43
- 2.6 Homomorphisms and Isomorphisms 46

Chapter 3. Integral Domains and Fields 52
- 3.1 Definition of Integral Domain 52
- 3.2 Ordered Integral Domains 54
- 3.3 Well-Ordering and Mathematical Induction 57

3.4 A Characterization of the Ring of Integers 62
3.5 Fields 65
3.6 Field of Quotients 67
3.7 A Few Properties of the Field of Rational Numbers 73
3.8 The Field of Real Numbers 75
3.9 Some Properties of the Field of Real Numbers 77
3.10 The Field of Complex Numbers 79
3.11 Geometric Representation and Trigonometric Form 83
3.12 The nth Roots of a Complex Number 88

Chapter 4. Factorization 92
4.1 Polynomial Rings 92
4.2 Divisors and the Division Algorithm 98
4.3 The Division Algorithm in Other Rings 102
4.4 Greatest Common Divisors 107
4.5 Greatest Common Divisors in Euclidean Domains 111
4.6 Unique Factorization in $F[x]$ 115
4.7 Roots of Polynomials 119
4.8 Unique Factorization in Euclidean Domains 126
4.9 Some Applications of the Fundamental Theorem 131
4.10 Rational Roots of Polynomials Over the Rational Field 134
*4.11 Integer Solutions of Linear Equations 136

Chapter 5. Field Extensions 140
5.1 Splitting Fields 140
5.2 Ideals 149
5.3 Ideals, Congruence Relations, and Homomorphisms 152
5.4 The Integers Modulo n 157
5.5 Finite Fields 163
5.6 Galois Theory of Finite Fields 169
5.7 Prime Polynomials Over the Rational Field 174
5.8 Polynomials Over the Real or Complex Numbers 178
*5.9 The Fundamental Theorem of Algebra 182

Chapter 6. Groups 186
6.1 Definitions and Simple Properties 186
6.2 Mappings and Permutation Groups 193
6.3 Homomorphisms and Isomorphisms 201
6.4 Cyclic Groups 207
6.5 Cosets and Lagrange's Theorem 213

6.6 The Symmetric Group S_n 217
6.7 Normal Subgroups and Factor Groups 224
*6.8 The 15-Puzzle 228

Chapter 7. Finite Groups 235
7.1 Groups Acting on Sets 235
7.2 Groups of Prime Power Order 240
7.3 Sylow's Theorem 242
7.4 Applications of Sylow's Theorem 247

Chapter 8. Finite Abelian Groups 250
8.1 Direct Sums of Subgroups 250
8.2 Cyclic Subgroups and Bases 254
8.3 Finite Abelian p-Groups 256
8.4 The Principal Theorems for Finite Abelian Groups 261
8.5 The Multiplicative Group of a Finite Field 266

Bibliography 268

Index 269

Preface

This edition, in common with the previous editions, is intended as a text for a *first* course in abstract algebra. The goal has continued to be that of making the exposition as simple and clear as possible, but sufficiently precise and thorough to furnish an honest introduction to the methods and results of abstract algebra.

In the present undergraduate curriculum, the first course in abstract algebra often serves several purposes. The student is expected to learn a collection of facts, to learn some methods and ideas, and, in addition, to learn to read and write correct mathematical proofs. All of these requirements have been carefully considered in the preparation of the text. The material in Chapter 4, for example, begins with the discussion of polynomial rings over a field, where certain basic facts are proved. The essential ideas are abstracted from this discussion and used to motivate the definition of a Euclidean ring. Some results are then proved in this more general context and then the application to the ring of integers is made. Along the way, the student is asked to provide proofs of some results, at first, by imitation of corresponding results for polynomial rings.

The important concept of factor rings is motivated by the desire to solve polynomial equations. The idea of finding a field extension in which a given equation has a solution is presented, the essential ideas are abstracted, and then applied to general rings. When this is applied to the special case of the ring of integers, we begin the study of finite fields. This study is carried far enough to prove the existence and uniqueness of fields of prime power order. This then provides the opportunity to introduce the student to Galois Theory in the setting of finite fields which is much less complicated than in the setting of general fields. It should be pointed out that this introduction of field theory is self-contained in that no linear algebra is used.

The three chapters on group theory carry the student through Sylow's Theorem and the structure theorem for finite abelian groups. Recognizing that there is far more material in this text than can reasonably be covered in many courses, we have made the final two chapters independent of each other. Either Chapter 7 or Chapter 8 may be covered after Chapter 6. Of course, there are other sections that may be treated as optional if time is short. A few of the final sections from Chapters 3, 4, 5, 7, or 8 could be omitted, depending on the needs of the class.

There is a considerable body of information contained in the examples and exercises. In particular, the reader will find frequent reference to rings and groups of two-by-two matrices which are introduced early as examples of noncommutative rings and groups. Even though no formal theory of matrices is presented in the text, the student should not have serious difficulty with this series of problems if the exercises on this topic are worked from the beginning. We have also included a number of topics that may be regarded as "enrichment" material. These topics would not ordinarily be covered but may be used for special projects, or as reading assignments for unusually motivated students. These sections are marked (*) in the Contents. The Fundamental Theorem of Algebra is one such topic; the proof given here can be read and understood by students with a good calculus background.

We are indebted to a considerable number of people in deciding what revisions would be desirable in this edition. In addition to a number of fairly short, but valuable, comments, rather extensive and constructive suggestions were made by Professors Donald Plank, Stockton State College; L. E. Sigler, Bucknell University; Henry Frandsen, University of Tennessee at Knoxville; James Dowdy, West Virginia University; F. D. Alexander, Stephen F. Austin State University; James R. Smith, Appalachian State University, for reading an essentially final version of the manuscript. Each of them made fairly detailed suggestions and also caught some slips, thus helping to improve the quality of the exposition at various points. Although not all suggestions were followed, they were all seriously considered and were of great value.

NHM
GJJ

Introduction to Modern Algebra

Introduction

The Logical Structure of Mathematics

The outstanding characteristic of modern abstract algebra, and indeed also of many other branches of modern mathematics, is its extensive use of what is known as the postulational or axiomatic method. The method itself is not new, since it was used by Euclid (about 300 B.C.) in his construction of geometry as a deductive science. However, in many ways the modern viewpoint is quite different from Euclid's, and the power of the method did not become apparent until the twentieth century.

At the present time, most students of mathematics have had some introduction to the postulational method before taking a course such as the one for which this book is designed as a text. However, in this brief preliminary chapter we present, primarily for review purposes, a brief outline of the basic ideas. Many illustrations will occur throughout this book.

Undefined Terms and Postulates

If one uses a dictionary to try to find the meaning of an unkown word, the definition is necessarily given in terms of another word (or words). If this word is also unknown, one must then try to find *its* meaning. A little thought will convince one that it is impossible for a dictionary to define all words; that is, a person must already know the meanings of some words or the dictionary is of no help.

In a mathematical system we avoid this problem by starting with a few *undefined terms* and make no attempt to define them. We then list some postulates or axioms (we shall use these words interchangeably) involving the undefined terms. The important fact is that the postulates tell us *all* that we need to know about the undefined terms.

As an example, in a study of plane geometry we might take *point*, *line*, *set*, and *between* as undefined terms.* Then among the postulates might be such as the following:

(1) A line is a set of points.
(2) There is exactly one line which contains two given distinct points.
(3) Given two distinct points on a line, there is another point on the line which is between them.

The first precise and modern treatment of plane geometry was due to Hilbert (1862–1943). In his approach, there were five undefined terms and fifteen postulates. We shall not pursue this example any further except to emphasize that it does not matter at all what one *thinks of* when the undefined terms are used—the important thing is that they have the properties given in the postulates.

Definitions

After one has the undefined terms, the postulates may be presented; in general, they describe assumed relationships between the undefined terms. Later on, it may be useful to define some additional terms. Of course, these definitions must depend only on the undefined terms and the postulates (or something which may have already been proved by use of the postulates). By way of illustration, in the above example of plane geometry we might define a new term ***segment*** as follows:

If A and B are distinct points on a line, the set of points consisting of the points A and B and all the points of the line between A and B is called a ***segment*** AB.

Note that postulates (1) and (3) assure us that there *are* some points on a segment. Observe, also, that this definition uses all the undefined terms mentioned above, namely, ***point***, ***line***, ***set***, ***between***.

Statements and Theorems

By a *statement* we shall mean a declaratory statement for which it is meaningful to say that it is either "true" or "false." We explicitly deny that a statement may be both true and false. Euclid considered the postulates to be statements which were obviously true in the physical world. The modern point of view is to consider the postulates of a system as statements to which the label *true* is assigned; that is, we are not directly concerned with physical reality, but the postulates are taken as true by the "rules of the game."

A statement which has been proved to be true in a given system is usually called a *theorem* (or a *proposition*). However, if the truth of a statement follows almost immediately from some theorem, the statement is usually called a *corollary* of the theorem. A result which is primarily obtained as an aid to the proof of a later theorem, and which may otherwise not be of great interest in itself, is often called a *lemma.*

---

* The word *set* may not be used explicitly in some treatments of geometry, but the concept underlies most of mathematics and we shall use it whenever we wish. This concept will be discussed briefly in the first section of the following chapter.

Implications

As an example of the next concept to be introduced, let ABC be a triangle in a given plane and let us consider the following statements:

(p) Triangle ABC is an equilateral triangle.
(q) Triangle ABC is an isosceles triangle.

Now consider the statement, "If p is true, then q is true." A statement of this form is called an *implication.* If it is a true implication, as in this example, we often say that *p implies q.* Many of the theorems in mathematics assert the truth of some implication. In proving a theorem of the form, "If p is true, then q is true," we call the statement p the *hypothesis* and the statement q the *conclusion.*

Suppose that p and q are given statements and that we are interested in the implication, "If p is true, then q is true." Associated with this implication in a natural way are three other implications as follows:

(Converse)	If q is true, then p is true.
(Inverse)	If p is false, then q is false.
(Contrapositive)	If q is false, then p is false.

We must emphasize that we are not asserting anything about whether these are true implications. In the example in which p and q are the statements made above about a triangle, the implication "If p is true, then q is true" is true. Neither the converse nor the inverse of this implication is true, but the contrapositive is readily seen to be true. This illustrates an important logical point as follows. It is always true that *an implication and its contrapositive are either both true or both false.* A little thought will prove this. Suppose that the implication "If p is true, then q is true" is true. Assume, now, that q is false. Then we must have that p is false, for if p were true, the given implication would show that q is true. Since q cannot be both true and false, we conclude that p must be false. A similar argument shows that if the contrapositive of a given implication is true, so is the given implication.

If r and s are statements such that both of the implications "If r is true, then s is true" and "If s is true, then r is true" are true, it is convenient to say that the statements r and s are *equivalent.* We have observed above that an implication and its contrapositive are equivalent. Thus, in order to prove some specified implication, we may just as well prove its contrapositive.

It may be observed that the contrapositive of the contrapositive of a given implication is the given implication. Moreover, the converse and the inverse of a given implication are contrapositives of each other.

Hypothesis and Conclusion

As we have indicated, a theorem is often of the form:

Theorem A. *If p is true, then q is true.*

Here we start with the truth of p as hypothesis, and prove the conclusion that q is true. However, we could equally well prove the contrapositive equivalent theorem:

Theorem A′. *If q is false, then p is false.*

In this case, we would take as hypothesis the fact that q is false, and the conclusion would be that p is false.

It may happen in a given situation that a proof of one of these theorems seems more natural than a proof of the other. In that case, we should of course always choose the more convenient form since the two theorems are equivalent.

Indirect Proofs

Sometimes an *indirect* proof is convenient. To give an indirect proof of Theorem A above, we would take as hypothesis that p is true and that q is false. The goal would then be to show that this hypothesis leads to a contradiction of one of the postulates or of something which has already been proved from the postulates. This contradiction would show that we cannot have simultaneously p true and q false, and this establishes the desired result.

"If and Only If"

It is not unusual to find a theorem which takes the following form:

Theorem B. *Statement p is true if and only if statement q is true.*

"If and only if" means that both of the following implications are true:

If q is true, then p is true.
If q is false, then p is false.

It will be observed that the contrapositive form of the second of these is

If p is true, then q is true.

Thus Theorem B asserts that p and q are equivalent statements. A theorem involving the words "if and only if" therefore always requires the proof of two implications.

Although we have here been speaking of implications, we may remark that a *definition* is always to be interpreted as an "if and only if" statement, although this particular grammatical construction is frequently not used in formulating a definition. Thus, the postulate

(1) A line is a set of points,

mentioned above, is by no means a definition of a line since it is not true that a set of points is necessarily a line. This postulate simply states *one* of the properties of a line. We have, in fact, suggested that *line* is frequently taken as one of the undefined terms in plane geometry.

This concludes our quick survey. Illustrations of all of the concepts introduced above will occur frequently in later chapters of this book.

Chapter 1

Some Fundamental Concepts

In this chapter we present a few basic concepts to be used repeatedly and introduce some convenient notation. Although the reader may very well have previously met some, or even all, of these concepts, they are so fundamental for our purposes that it seems desirable to present them here in some detail.

1.1 SETS

The concept of *set* (class, collection, aggregate) is fundamental in mathematics as it is in everyday life. A related concept is that of *element* of a set. We make no attempt to define these terms; that is, we shall consider them to be undefined terms in our system. However, we shall presently give some examples that will illustrate the sense in which they are being used.

First of all, we may say that a set is made up of elements. In order to give an example of a set we need, therefore, to exhibit its elements or to give some rule that will specify its elements. We shall often find it convenient to denote sets by capital letters and elements of sets by lower-case letters. If a is an element of the set A, we may indicate this fact by writing $a \in A$ (read, "a is an element of A"). Also, $a \notin A$ will mean that a is not an element of the set A. If both a and b are elements of the set A, we may write $a, b \in A$.

If P is the set of all positive integers, $a \in P$ means merely that a is a positive integer. Certainly, then, it is true that $1 \in P$, $2 \in P$, and so on. If B is the set of all triangles in a given plane, $a \in B$ means that a is one of the triangles in this plane. If C is the set of all books in the Library of Congress, then $a \in C$ means that a is one of these books. We shall presently give other examples of sets.

If $a, b \in A$ and we write $a = b$, *it is always to be understood that these are identical elements of* A. In other words, a and b are merely different symbols designating the same element of A. If $a, b \in A$ and it is not true that $a = b$, we may indicate this fact by writing $a \neq b$ and may say that a and b are *distinct* elements of A.

If A and B are sets with the property that every element of A is also an element of B, we call A a *subset* of B and write $A \subseteq B$ (read, "A is contained in B"). An alternative way of expressing the fact that $A \subseteq B$ is to write $B \supseteq A$ (read, "B contains A"). If it is not true that $A \subseteq B$, we may indicate this fact by writing $A \nsubseteq B$. We should point out that for every set A it is true that $A \subseteq A$ and hence, according to our definition, one of the subsets of A is A itself.

If $A \subseteq B$ and also $B \subseteq A$, then A and B have exactly the same elements and we say that these sets are *equal* and indicate this by writing $A = B$. If it is not true that $A = B$, we may write $A \neq B$. If $A \subseteq B$ and $A \neq B$, then we say that A is a *proper subset* of B and indicate this fact by the notation $A \subset B$ (read, "A is properly contained in B"). Clearly, $A \subset B$ means that every element of A is an element of B and, moreover, B contains at least one element which is not an element of A.

Sometimes, as has been the case so far, we may specify a set by stating in words just what its elements are. Another way of specifying a set is to exhibit its elements, usually enclosed between braces. Thus, $\{x\}$ indicates the set which consists of the single element x, $\{x, y\}$ the set consisting of the two elements x and y, and so on. We may write $A = \{1, 2, 3, 4\}$ to mean that A is the set whose elements are the positive integers 1, 2, 3, and 4. If P is the set of all positive integers, by writing

$$K = \{a \mid a \in P, a \text{ divisible by } 2\}$$

we mean that K consists of all elements a having the properties indicated after the vertical bar. That is, a is a positive integer and is divisible by 2. Hence, K is just the set of all *even* positive integers. We may also write

$$K = \{2, 4, 6, 8, \ldots\},$$

the dots indicating that all even positive integers are included in this set. As another example, if

$$D = \{a \mid a \in P, a < 6\},$$

then $D = \{1, 2, 3, 4, 5\}$.

Whenever we specify a set by exhibiting its elements, it is to be understood that the indicated elements are distinct. Thus, for example, if we write $B = \{x, y, z\}$, we mean to imply that $x \neq y$, $x \neq z$, and $y \neq z$.

For many purposes, we allow for the possibility that a set may have no elements. This set with no elements we shall call the *empty set*. According to the definition of subset given above, the empty set is a subset of every set.

Moreover, it is a proper subset of every set except the empty set itself. The empty set is often designated by $\varnothing$, and thus we have $\varnothing \subseteq A$ for every set A.

If A and B are sets, the elements that are in both A and B form a set called the *intersection* of A and B, denoted by $A \cap B$. Of course, if A and B share no elements, $A \cap B = \varnothing$.

If A and B are sets, the set consisting of those elements which are elements either of A or of B (or of both) is a set called the *union* of A and B, denoted by $A \cup B$.

As examples of the concepts of intersection and union, let $A = \{1,2,3\}$, $B = \{2,4,5\}$, and $C = \{1,3,6\}$. Then we have $A \cap B = \{2\}$, $A \cap C = \{1,3\}$, $B \cap C = \varnothing$, $A \cup B = \{1,2,3,4,5\}$, $A \cup C = \{1,2,3,6\}$, and $B \cup C = \{1,2,3,4,5,6\}$.

Although we have defined the intersection and the union of only *two* sets, it is easy to extend these definitions to any number of sets, as follows. The *intersection* of any number of given sets is the set consisting of those elements which are in all the given sets and the *union* is the set consisting of those elements which are in at least one of the given sets.

If A, B, and C are sets, each of the following is an almost immediate consequence of the various definitions which we have made:

$A \cap B \subseteq A$ and $A \cap B \subseteq B$.
$A \subseteq A \cup B$ and $B \subseteq A \cup B$.
$A \cap B = A$ if and only if $A \subseteq B$.
$A \cup B = A$ if and only if $B \subseteq A$.
If $B \subseteq C$, then $A \cup B \subseteq A \cup C$ and $A \cap B \subseteq A \cap C$.

In two of these statements, we have used the expression "if and only if." Thus, in accordance with the explanation given in the preceding chapter, we have to establish two different implications. For example, to show that "$A \cap B = A$ if and only if $A \subseteq B$," we need to verify the following implications:

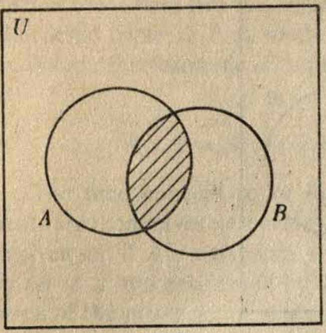
If $A \cap B = A$, then $A \subseteq B$.
If $A \subseteq B$, then $A \cap B = A$.

Naturally, we could just as well use the contrapositive form of either or both of these implications. For example, the second one is equivalent to the implication

If $A \cap B \neq A$, then $A \not\subseteq B$.

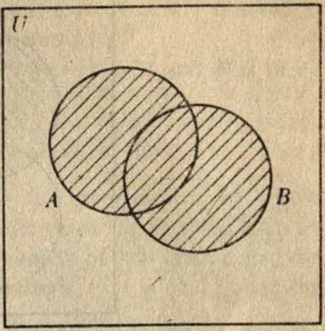
We leave to the reader the simple verification of all the implications stated above.

In working with sets, so-called Venn diagrams are sometimes used to give a purely symbolic, but convenient, geometric indication of the relationships involved. Suppose, for the moment, that all sets being considered are



$A \cap B$

Figure 1



$A \cup B$

Figure 2

subsets of some fixed set U. In Figures 1 and 2, the points within the square represent elements of U. If A and B are subsets of U, then the elements of A and B may be represented by the points within indicated circles (or any other closed regions). The intersection and the union of the sets A and B are then represented by the shaded regions in Figures 1 and 2, respectively.

Of course, the use of a Venn diagram is not meant to imply anything about the nature of the sets being considered (whether or not indicated intersections are nonempty, and so on). Moreover, such a diagram cannot in itself constitute a proof of any fact, but it may be quite helpful in suggesting a proof.

Let us make the following remarks by way of emphasis. A frequent problem is that of proving the equality of two sets. Suppose that C and D are given sets and it is required to prove that $C = D$. By definition of equality of sets, we need to show that $C \subseteq D$ and $D \subseteq C$. Sometimes one or both of these conditions follow easily from given facts. If not, the standard procedure is to start with an arbitrary element of C and show that it is an element of D, and then do the same thing with C and D interchanged. When we write "let $x \in C$" or "if $x \in C$," we mean that x is to represent a completely arbitrary element of the set C. Hence, to show that $C \subseteq D$, we only need to show that "if $x \in C$, then $x \in D$." Of course, any other symbol could be used in place of x. Let us now give an example by way of illustration.

Example: If A, B, and C are sets, prove that

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

Solution: First, let us take advantage of the opportunity to give another illustration of a Venn diagram. If we think of the meaning of $A \cup (B \cap C)$ as consisting of all elements of A together with all elements that are in both B and C, we see that the set $A \cup (B \cap C)$ may be represented by the shaded portion of the Venn diagram in

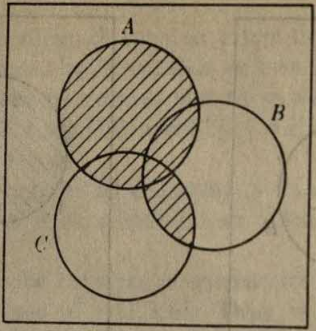


Figure 3

Figure 3. We leave it to the reader to verify that this same shaded region also represents the set $(A \cup B) \cap (A \cup C)$.

We now proceed to give a formal proof of the required formula. Clearly, $B \cap C \subseteq B$, so $A \cup (B \cap C) \subseteq A \cup B$. Similarly, $B \cap C \subseteq C$, and hence $A \cup (B \cap C) \subseteq A \cup C$. It follows that

$$A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C),$$

and we have obtained inclusion one way. To obtain inclusion the other way, let $x \in (A \cup B) \cap (A \cup C)$ and let us show that $x \in A \cup (B \cap C)$. Now $x \in A \cup B$ and also $x \in A \cup C$. If $x \in A$, then surely $x \in A \cup (B \cap C)$. If $x \notin A$, then $x \in B$ and $x \in C$, so that $x \in B \cap C$, and again we have that $x \in A \cup (B \cap C)$. This shows that $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$, and the proof is therefore complete.

If A and B are subsets of some set U, it is at times convenient to have a notation for the set of elements of A that are *not* elements of B. This set is sometimes denoted by $A - B$, but we shall use the notation $A \setminus B$ in order that there can be no confusion with other uses of the "minus" symbol. For example, if $A = \{1, 2, 3\}$ and $B = \{2, 4\}$, then $A \setminus B = \{1, 3\}$.

Another important concept is illustrated by the familiar idea of coordinates of a point in a plane. A point is determined by an ordered pair (x, y) of real numbers. The word *ordered* is meant to imply that the order of writing the two numbers x and y is important, that is, that (x, y) is to be considered as a different pair than (y, x) unless, of course, x and y happen to be equal real numbers. If **R** denotes the set of all real numbers, the set of all ordered pairs of elements of **R** is frequently called the *Cartesian product* of **R** by **R** and designated by **R** × **R**. More generally, if A and B are any sets, the set of all ordered pairs (a, b), where $a \in A$ and $b \in B$, is the *Cartesian product* of A by B, designated by $A \times B$. It may happen, of course, that A and B are identical sets, as in the illustration given above. It is obvious how to define the Cartesian product of more than two sets. Thus, for example, the set $A \times B \times C$ is the set of all ordered triples (a, b, c), where $a \in A$, $b \in B$, and $c \in C$.

As another example of a product set, if $A = \{1, 2, 3\}$ and $B = \{u, v\}$, then

$$A \times B = \{(1, u), (1, v), (2, u), (2, v), (3, u), (3, v)\}.$$

The final concept to be introduced in this section is a set whose elements are themselves sets; more specifically, the elements are the subsets of some given set. If A is a given set, the set of all subsets of A is often called the *power set* of A and designated by $\mathscr{P}(A)$. For example, if $A = \{1, 2\}$, then the elements of the power set of A are the subsets of A; that is,

$$\mathscr{P}(A) = \{\varnothing, \{1\}, \{2\}, \{1, 2\}\}.$$

EXERCISES

1. If $A = \{a, b, c\}$, $B = \{c, x, y\}$, and $C = \{x, y\}$, determine each of the following sets: $A \cap B, A \cap C, A \cup B, A \cup C, A \times C, C \times A, A \backslash B, \mathscr{P}(A), C \times \mathscr{P}(C)$.

2. Let P be the set of all positive integers, and define subsets of P as follows:

$$F = \{a \mid a \in P, a < 10\},$$
$$G = \{a \mid a \in P, a > 5\},$$
$$H = \{a \mid a \in P, a \text{ divisible by } 3\}.$$

Determine each of the following sets: $F \cap G, F \cap H, G \cap H, F \cup G, F \cup H, G \cup H$.

3. If A, B, and C are sets, draw Venn diagrams to illustrate and then give a formal proof that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

4. If A, B, and C are subsets of some set U, prove each of the following:

(i) $A \backslash (B \cup C) = (A \backslash B) \cap (A \backslash C)$,

(ii) $A \backslash (B \cap C) = (A \backslash B) \cup (A \backslash C)$.

5. If k is a positive integer, show that a set with $k + 1$ elements has twice as many subsets as a set with k elements.

6. Use the result of the preceding exercise to give a plausible reason (not necessarily a formal proof) why the following statement is true: If A is a set with n elements, where n is a positive integer, then the set $\mathscr{P}(A)$ has 2^n elements.

7. The coefficient of $x^r y^{n-r}$ $(0 \leq r \leq n)$ in the binomial expansion of $(x + y)^n$ is the number of ways in which r objects can be chosen from a set of n objects. Use this fact, applied to $(1 + 1)^n$, to give a proof of the result stated in the preceding exercise.

8. If the set A has n elements, how many elements are there in the set $\mathscr{P}(\mathscr{P}(A))$?

9. If the set A has n elements, n a positive integer, prove that A has as many subsets with an even number of elements as it has subsets with an odd number of elements.

1.2 MAPPINGS

As a first illustration of the concept to be introduced in this section, let C be the set of all books in the Library of Congress and P the set of all positive integers. Corresponding to each book there is a unique positive integer, namely, the number of pages in the book. That is, to each element of C there corresponds in this way a unique element of P. This is an example of a mapping of the set C into the set P. As another illustration, let N be the set of all names occurring in a given telephone directory, and L the set of the twenty-six letters of the alphabet. We may then associate with each name the first letter of the surname, and this then defines a mapping of N into L. Additional examples will be given after the following definition.

1.1 Definition. A ***mapping*** of a set A into a set B is a correspondence that associates with each element a of A a unique element b of B. The notation $a \to b$ is sometimes used to indicate that b is the element of B that is associated with the element a of A under a given mapping. We may say that ***a maps into b*** or that b is the ***image*** of a under this mapping.

In order to avoid some trivial special cases, whenever we consider a mapping of a set A into a set B we shall always assume that the sets A and B are not empty.

Let us now give an example of a mapping of the set $S = \{1, 2, 3, 4\}$ into the set $T = \{x, y, z\}$. To specify such a mapping, we select an element of T to be the image of each element of S. Thus

1.2 $$1 \to x, \quad 2 \to y, \quad 3 \to x, \quad 4 \to y$$

defines a mapping of S into T in which x is the image of 1, y the image of 2, and so on. Note that although every element of S is required to have an image in T, it need not be true that every element of T occurs as the image of at least one element of S.

Before proceeding, let us observe that a *function*, as the term is often used, is just a mapping of the set **R** of all real numbers (or of some subset of **R**) into the same set **R**. For example, the function f defined by $f(x) = x^2 + x + 1$ is the mapping $x \to x^2 + x + 1$ which associates with each real number x the real number $x^2 + x + 1$. In this setting, the mapping is denoted by f and the image of the number x under the mapping f by $f(x)$.

Although we are now concerned with arbitrary sets (not just sets of real numbers), the function notation of the preceding paragraph will be used. Mappings will henceforth be denoted by Greek letters, such as $\alpha, \beta, \gamma, \ldots$. If α is a mapping of A into B, the image of the element a of A will be denoted by $\alpha(a)$ or, occasionally, by just αa. For example, let β be the mapping 1.2 of S into T.

Then, instead of writing 1.2, we may write

1.3 $$\beta(1) = x, \quad \beta(2) = y, \quad \beta(3) = x, \quad \beta(4) = y.$$

Another mapping γ of S into T is defined by

1.4 $$\gamma(1) = x, \quad \gamma(2) = y, \quad \gamma(3) = y, \quad \gamma(4) = z.$$

We shall presently use these mappings to illustrate certain additional concepts.

It is customary to write $\alpha: A \to B$ to indicate that α is a mapping of the set A into the set B. We may also sometimes write $a \to \alpha(a)$, $a \in A$, to indicate this mapping, it being understood that for each $a \in A$, $\alpha(a)$ is a uniquely determined element of B. If we have mappings $\alpha: A \to B$ and $\beta: A \to B$, we naturally consider these mappings to be equal and write $\alpha = \beta$, if and only if $\alpha(a) = \beta(a)$ for every $a \in A$. Thus, for the mappings $\beta: S \to T$ and $\gamma: S \to T$ exhibited above, we have $\beta \neq \gamma$ since, for example, $\beta(3) = x$ and $\gamma(3) = y$.

We may point out that associated with the mapping 1.3 in a natural way is a unique subset of the product set $S \times T$—namely, the set $\{(1, x), (2, y), (3, x), (4, y)\}$, in which we use, as the second element of a pair, the image of the first element under the mapping β. More generally, if $\alpha: A \to B$ is a mapping of A into B, we may associate with the mapping α the subset of $A \times B$ whose elements are the pairs $(a, \alpha(a))$, $a \in A$. Conversely, suppose that T is a subset of $A \times B$ with the following two properties:

(1) For each element a of A, there exists an element of T of the form (a, b); that is, each element a of A is the first element of some ordered pair in T.
(2) If $(a, b) \in T$ and $(a, c) \in T$, then $b = c$.

These two properties merely assure us that if $a \in A$, there exists exactly *one* element b of B such that $(a, b) \in T$. Since the element b of B is uniquely determined in this way by a, we may define a mapping $\alpha: A \to B$ by defining $\alpha(a) = b$. For this reason, a mapping of A into B is sometimes *defined* to be a subset of $A \times B$ with the two properties stated above. Although we shall not take this viewpoint, note that this approach makes it possible to define a mapping without using the undefined word *correspondence* which was used in Definition 1.1.

In studying mappings it is sometimes suggestive to make use of a geometrical diagram. Figure 4 suggests that α is a mapping of A into B and that under this mapping each element a of A has image $\alpha(a)$ in B.

In the particular mapping $\beta: S \to T$ given by 1.3, the element z of T does not occur as the image of any element of S. However, in the mapping $\gamma: S \to T$, defined in 1.4, every element of T is the image of at least one element of S. The language for stating this essential difference between these mappings is given in the following definition.

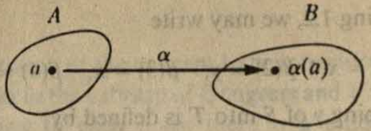


Figure 4

1.5 Definition. A mapping α of A into B is said to be a mapping of A ***onto B*** if and only if every element of B is the image of at least one element of A under the mapping α.

Thus, γ is a mapping of S onto T, whereas β is not a mapping of S onto T. It is important to observe that "into" is *not* the opposite of "onto." According to our language, every mapping is a mapping of some set into some set. That is, "onto" is a special case of "into," and if $\alpha: A \to B$ is a mapping of A *onto* B, it is perfectly correct to say that it is a mapping of A into B (although this doesn't give the maximum amount of available information).

If $\alpha: A \to B$ is a mapping of A into B, let us denote by $\alpha(A)$ or αA the set of all elements of B that occur as images of elements of A under the mapping α, that is,

$$\alpha(A) = \{\alpha(a) \mid a \in A\}.$$

Thus α is a mapping of A onto B if and only if $\alpha A = B$. In any case, an arbitrary mapping $\alpha: A \to B$ may be considered as defining a mapping of A *onto* the subset $\alpha(A)$ of B. Thus, associated with each mapping is an onto mapping if we suitably restrict the set in which the images lie.

One additional concept plays an important role in the study of mappings. In the mapping $\beta: S \to T$, defined by 1.3, we see that both 1 and 3 have x as image. Similarly, $\gamma: S \to T$, defined by 1.4, is such that both 2 and 3 have y as image. Now let $T = \{x, y, z\}$ as above, and let $U = \{r, s, t, u\}$. Then the mapping $\theta: T \to U$ defined by

1.6 $$\theta(x) = t, \quad \theta(y) = r, \quad \theta(z) = u$$

is such that every element of U which occurs as an image of some element of T is the image of exactly one element of T. This property has a name which we proceed to introduce.

1.7 Definition. A mapping $\alpha: A \to B$ is said to be a ***one-one mapping*** of A into B if and only if distinct elements of A have distinct images in B; equivalently, if $a_1, a_2 \in A$ such that $\alpha(a_1) = \alpha(a_2)$, then $a_1 = a_2$.

The mapping $\theta: T \to U$ defined by 1.6 is an example of a one-one mapping. Note, however, that it is not an onto mapping. Hence a one-one mapping may or may not be an onto mapping. Clearly, also, an onto mapping need not be a one-one mapping.*

We now give some additional examples to illustrate these concepts.

Example 1: Let C be a nonempty subset of the set D. The mapping $\phi:C \to D$ defined by $\phi(c) = c$ for each $c \in C$ is a one-one mapping of C into D. It is an onto mapping if and only if $C = D$.

Example 2: Let $\alpha:A \times B \to A$ be defined by $\alpha(a, b) = a$ for each $(a, b) \in A \times B$. This is certainly an onto mapping. However, if $b_1, b_2 \in B$ with $b_1 \neq b_2$, then $\alpha(a, b_1) = \alpha(a, b_2)$ with $(a, b_1) \neq (a, b_2)$, so the mapping is not a one-one mapping. It will be a one-one mapping if and only if B has exactly one element. The mapping α of this example is called the *projection* of $A \times B$ onto A. Similarly, one can define the projection of $A \times B$ onto B.

Example 3: Let $\mathbf{Z}$ be the set of all integers and $\alpha:\mathbf{Z} \to \mathbf{Z}$ be defined by $\alpha(i) = 2i + 1, i \in \mathbf{Z}$. In contrast to most of our previous examples, this is an example of a mapping of the set $\mathbf{Z}$ into the same set $\mathbf{Z}$. To determine whether α is an onto mapping, let j be an arbitrary element of $\mathbf{Z}$ and let us find whether j is the image of some element. That is, we need to determine whether there exists an integer i such that $\alpha(i) = 2i + 1 = j$. Clearly, there will be no such integer i if j is even since $2i + 1$ is odd for every integer i. Thus, α is not an onto mapping. Is it a one-one mapping? To answer this question, suppose that $i_1, i_2 \in \mathbf{Z}$ such that $\alpha(i_1) = \alpha(i_2)$, that is, such that $2i_1 + 1 = 2i_2 + 1$. It follows that $i_1 = i_2$ and α is therefore a one-one mapping.

Identity Map

The mapping of a set A into itself in which each element is its own image is often called the *identity mapping* on the set A. If we denote this identity mapping by ϵ_A, we see that the identity mapping ϵ_A on A is defined by $\epsilon_A(a) = a$ for each $a \in A$. Thus, the mapping ϕ of Example 1 is the identity mapping on C if and only if $C = D$. It is clear that an identity mapping is always one-one and onto.

Inverse Map

Suppose that $\alpha:A \to B$ is a *one-one* mapping of A *onto* B, that is, it has both the "one-one" and "onto" properties. Then each element of B can be expressed in the form $\alpha(a)$ for exactly one element a of A. We can therefore define in a natural way a mapping of B onto A by making a the image of $\alpha(a)$ for each $a \in A$. This particular mapping of B onto A is often denoted by α^{-1} since it reverses the effect of α. That is, the mapping $\alpha^{-1}:B \to A$ is defined by $\alpha^{-1}(\alpha a) = a, a \in A$. Clearly, α^{-1} is a one-one mapping of B onto A. The simple relationship between the mappings α and α^{-1} may possibly be suggested by the

---

* Some other terms will be found in the literature as follows. An onto mapping is also called a *surjection;* a one-one mapping may be called an *injection*. A mapping which is both one-one and onto is also called a *bijection*.

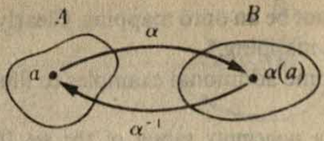


Figure 5

diagram of Figure 5. The reader should carefully observe that the mapping α^{-1}, as here defined, exists if and only if α is a one-one mapping of A onto B.

Clearly, a one-one mapping of A onto B may be thought of as a pairing of the elements of A and the elements of B. In view of the mutual relationship between A and B, a one-one mapping of A onto B (or of B onto A) is sometimes called a *one-to-one correspondence between A and B.*

It will probably seem reasonable (and it is indeed true although we shall not discuss this fact here) that if set C has n elements for some positive integer n, then there will exist a one-one mapping of C onto D if and only if D also has n elements.

EXERCISES

1. Let **Z** be the set of all integers, and $i \in \mathbf{Z}$. Determine in each case whether the indicated mapping α of **Z** into **Z** is an onto mapping and whether it is a one-one mapping.

(a) $\alpha(i) = i + 3$, (b) $\alpha(i) = i^2 + i$,
(c) $\alpha(i) = i^3$, (d) $\alpha(i) = 2i - 1$,
(e) $\alpha(i) = -i + 5$, (f) $\alpha(i) = i - 4$.

2. Let **R** be the set of all real numbers, and $x \in \mathbf{R}$. Determine in each case whether the indicated mapping α of **R** into **R** is an onto mapping and whether it is a one-one mapping.

(a) $\alpha(x) = 2x + 1$, (b) $\alpha(x) = 1 - x$,
(c) $\alpha(x) = x^2$, (d) $\alpha(x) = x^3$,
(e) $\alpha(x) = x^2 + x$, (f) $\alpha(x) = 4x$,

(g) $\alpha(x) = \begin{cases} x \text{ if } x \text{ is rational,} \\ 2x \text{ if } x \text{ is irrational.} \end{cases}$

3. Let P be the set of all positive integers, and $n \in P$. Determine in each case whether the indicated mapping of P into P is an onto mapping and whether it is a one-one mapping.

(a) $\alpha(n) = 2n$, (b) $\alpha(n) = n + 1$,
(c) $\alpha(n) = n^2$. (d) $\alpha(n) = 1, \alpha(n) = n - 1$ for $n > 1$.

4. Give an example of a mapping of the set P of all positive integers into the set P such that every element of P is the image of exactly two elements.

5. If P is as in the preceding exercise, give several examples of mappings $\alpha: P \to P$, other than the identity mapping on P, such that α is a one-one and onto mapping.

6. If **R** is the set of all real numbers, use the fact that every cubic equation with real coefficients has a real root to show that the mapping α of **R** into **R** defined by $\alpha(x) = x^3 - x, x \in \mathbf{R}$, is a mapping of **R** onto **R**. Is it a one-one mapping?

7. If **R** is the set of all real numbers, why doesn't the formula $\alpha(x) = 1/x, x \in \mathbf{R}$, define a mapping of **R** into **R**?

8. If $A = \{1, 2, 3\}$ and $B = \{x, y\}$, verify that there exist eight mappings of A into B and nine mappings of B into A. How many mappings of A onto B are there?

9. Let A be a set with m elements and B a set with n elements (m and n positive integers). Formal proofs are not required, but in each of the following give some indication why you believe your conclusion to be correct.

   (a) Determine the number of mappings of A into B.
   (b) If $n \geq m$, determine the number of one-one mappings of A into B.
   (c) If $m = n$, determine the number of one-one mappings of A onto B.

10. Let **R** be the set of all real numbers, and let $x \in \mathbf{R}$. Use any calculus methods which you know to determine whether each of the following mappings α of **R** into **R** is an onto mapping and whether it is a one-one mapping:

    (a) $\alpha(x) = e^x$,
    (b) $\alpha(x) = \sin x$,
    (c) $\alpha(x) = x + \sin x$,
    (d) $\alpha(x) = \frac{1}{2}x + \sin x$.

1.3 EQUIVALENCE RELATIONS

As a simple illustration of the next concept to be introduced, let the "less than" symbol "$<$" have the usual meaning as applied to integers. Since for every ordered pair (i, j) of integers, $i < j$ is either true or false, we say that "$<$" is a relation *defined* on the set **Z** of integers.

In general, let A be a given set. We then say that a *relation* R is defined on A if for each ordered pair (a, b) of elements of A it is true or false that a is in the relation R to b. It is to be understood that if a and b are given, enough information is available to determine whether or not a is in the relation R to b. It is customary to write aRb to indicate that a is in the relation R to b. If, for example, we let R be the relation $<$ on **Z**, then iRj merely means that $i < j$.

If R is a given relation on a set A, associated with this relation in a uniquely determined manner is a subset of the product set $A \times A$, namely, the subset consisting of all ordered pairs (a, b) such that aRb. Conversely, given a subset R of $A \times A$, we can use it to determine a relation R on A by simply defining aRb to mean that the ordered pair (a, b) is an element of R. For this reason, a relation on a set A is sometimes *defined* to be a subset of $A \times A$.

We shall usually not be concerned with relations in general, but primarily with those relations having the particular properties stated in the following definition.

1.8 Definition. A relation R defined on a set A is called an ***equivalence relation*** if it has the following three properties, where a, b, and c are arbitrary elements of A:

(1) aRa (*reflexive property*).
(2) If aRb, then bRa (*symmetric property*).
(3) If aRb and bRc, then aRc (*transitive property*).

In the future, we shall usually use "$\sim$" to denote an equivalence relation. It may then be convenient to read $a \sim b$ as "a is equivalent to b." We shall sometimes write $a \nsim b$ to indicate that a is not equivalent to b.

Let us emphasize how property (1) above differs in an essential way from the other two properties. If we have an equivalence relation "$\sim$" defined on a set A, then property (1) asserts that $a \sim a$ for *every* element a of A. On the other hand, if a and b are given, property (2) says nothing about whether it is true that $a \sim b$, only that *if* it is true, then we must also have $b \sim a$. A similar remark holds for the transitive property. In other words, both the symmetric property and the transitive property assert the truth of an implication.

The relation "$<$" on the set **Z** of all integers is not an equivalence relation since it has neither the reflexive property nor the symmetric property. The relation "$\leq$" has the reflexive property and the transitive property, but not the symmetric property. Of course, "$=$" is an equivalence relation on **Z** (or on any other set), as is perhaps suggested by the word "equivalence."

We shall now give a few examples of equivalence relations, but many more will occur in later chapters of this book. Let T be the set of all triangles in a fixed plane, and let a and b be arbitrary elements of T. Then "$\sim$" is an equivalence relation on T if we agree to define "$\sim$" in any one of the following ways:

(i) $a \sim b$ to mean "a is congruent to b,"
(ii) $a \sim b$ to mean "a is similar to b,"
(iii) $a \sim b$ to mean "a has the same area as b,"
(iv) $a \sim b$ to mean "a has the same perimeter as b."

As another example of an equivalence relation, let **Z** be the set of all integers, and let us define $a \equiv b$ to mean that $a - b$ has 3 as a factor, that is, that there exists an integer n such that $a - b = 3n$. It is then readily verified that "$\equiv$" has the three defining properties of an equivalence relation. Furthermore, every integer is equivalent to one of the three integers 0, 1, 2. In this connection, consider the following three subsets of **Z**:

$$\begin{aligned} J &= \{\ldots, -9, -6, -3, 0, 3, 6, 9, \ldots\}, \\ K &= \{\ldots, -8, -5, -2, 1, 4, 7, 10, \ldots\}, \\ L &= \{\ldots, -7, -4, -1, 2, 5, 8, 11, \ldots\}. \end{aligned}$$

It will be observed that every integer is in exactly one of these subsets. In other words, the union of these three subsets is **Z** and the intersection of any two of them is the empty set. Moreover, J can be characterized as the set of all elements of **Z** that are equivalent to 0 (or to any other element of J), and similar characterizations can be given for K and L. The sets J, K, and L are examples of a concept which we proceed to define.

1.9 Definition. Let A be a set and "$\sim$" an equivalence relation defined on A. If $a \in A$, the subset of A which consists of all elements x of A such that $x \sim a$ is called an ***equivalence class***. This equivalence class will frequently be denoted by $[a]$.

This definition of the equivalence class $[a]$ can be written formally as follows:

1.10
$$[a] = \{x \mid x \in A, x \sim a\}.$$

In the above example, note that $J = [0]$, $K = [1]$, and $L = [2]$; also that $[0] = [3] = [6]$, and so on. Hence there are just the three different equivalence classes.

To return to the general definition, let us consider a few properties of equivalence classes. First, since $a \sim a$ by the reflexive property of an equivalence relation, we always have $a \in [a]$; that is, $[a]$ is the equivalence class which contains a. This shows that every element of A is in at least one equivalence class. Other important properties of equivalence classes are the following, where a and b are elements of the set A:

1.11 (i) $[a] = [b]$ if and only if $a \sim b$;
(ii) if $[a] \cap [b] \neq \varnothing$, then $[a] = [b]$.

As a first step in proving 1.11(i), let us assume that $[a] = [b]$ and show that $a \sim b$. It has been pointed out that $a \in [a]$, and hence we have $a \in [b]$. By definition of the equivalence class $[b]$, it follows that $a \sim b$, as we wished to show. Conversely, let us now assume that $a \sim b$. If $x \in [a]$, then $x \sim a$ by definition of $[a]$. Now we have $x \sim a$ and $a \sim b$, so the transitive property of an equivalence relation assures us that $x \sim b$. This then implies that $x \in [b]$, and we have therefore proved that $[a] \subseteq [b]$. We leave as an exercise the similar proof that $[b] \subseteq [a]$, from which we conclude that $[a] = [b]$, as desired.

We now prove 1.11(ii). Since $[a] \cap [b] \neq \varnothing$, there exists at least one element s of A such that $s \in [a]$ and also $s \in [b]$. It follows that $s \sim a$ and $s \sim b$. (Why?) By the symmetric property of an equivalence relation, we have $a \sim s$. Since $a \sim s$ and $s \sim b$, the transitive property assures us that $a \sim b$. The fact that $[a] = [b]$ the follows at once from 1.11(i).

Partitions

A collection of nonempty subsets of a set A is often called a *partition* of A if A is the union of these subsets and any two of the subsets have empty intersection. In view of 1.11(ii), together with the fact that every element of A is in some equivalence class, we see that the different equivalence classes relative to an equivalence relation defined on A form a partition of A. Conversely, suppose that we are given a partition of the set A. If $a, b \in A$, and we define "$\sim$" by asserting that $a \sim b$ if and only if a and b are elements of the same subset of A in the given partition, then it is easy to verify that "$\sim$" is an equivalence relation; clearly the equivalence classes are the different subsets of A in the given partition. It follows that there is a one-to-one correspondence between equivalence relations on A and partitions of A.

1.4 OPERATIONS

There is one other term that we wish to introduce in this preliminary chapter. First, we consider a familiar concept as follows. Let $\mathbf{Z}$ be the set of all integers. Associated with each ordered pair (i, j) of elements of $\mathbf{Z}$ there is a uniquely determined element $i + j$ of $\mathbf{Z}$. Accordingly, we say that addition, denoted by "$+$," is an operation on $\mathbf{Z}$. More precisely, we may call it a *binary* operation to emphasize that it is defined for each ordered *pair* of elements of $\mathbf{Z}$. The general definition is as follows.

> **1.12 Definition.*** Let A be a given set. ***A binary operation*** "$\circ$" on A is a correspondence that associates with each ordered pair (a, b) of elements of A a uniquely determined element $a \circ b$ of A.

It may be emphasized that if a binary operation "$\circ$" is defined on the set A, this is supposed to imply that $a \circ b$ must always be an element *of the set* A, for $a, b \in A$. In this connection, suppose that B is a subset of A and that we have an operation "$\circ$" defined on A. If it happens that for $x, y \in B$ it is always true that $x \circ y$ (which we know is an element of A) is actually an element of the subset B, then it is customary to say that B is *closed* under (or with respect to) the operation "$\circ$." Of course, in this case, we may consider "$\circ$" to be an operation on the set B. As examples, consider ordinary addition and multiplication defined on the set $\mathbf{Z}$ of all integers, and let S be the set of all *odd* integers. Then we see that S is closed under multiplication but is not closed under addition. The set of all *even* integers is closed under both addition and multiplication.

---

* Expressed more formally, this definition merely asserts that a binary operation "$\circ$" on A is a mapping of $A \times A$ into A. The image of an element (a, b) of $A \times A$ under this mapping is then denoted by $a \circ b$.

Later on we shall seldom have occasion to use any unfamiliar symbol to denote an operation. For the most part, we shall find it convenient to call an operation "addition" or "multiplication," and to use the familiar notations $a + b$ and $a \cdot b$ (or simply ab). However, for the moment we continue to use the symbol "$\circ$" for a binary operation on a set A. We may emphasize that saying that "$\circ$" is a binary operation on A asserts that $a \circ b$ is a uniquely determined element of A for *every* $a \in A$ and *every* $b \in A$. Some important concepts are introduced in the following definition.

1.13 Definition. Let "$\circ$" be a binary operation defined on the set A. Then

(i) The operation "$\circ$" is said to be a ***commutative*** operation if and only if $a \circ b = b \circ a$ for all $a, b \in A$.
(ii) The operation "$\circ$" is said to be an ***associative*** operation if and only if $(a \circ b) \circ c = a \circ (b \circ c)$ for all $a, b, c \in A$.
(iii) An element e of A is said to be an ***identity*** for the operation "$\circ$" if and only if $a \circ e = e \circ a = a$ for every $a \in A$.

As examples of these concepts, let us again consider the set **Z** of all integers. For the present, we assume as known the familiar properties of addition and multiplication on **Z**, in particular, that they are both commutative and associative. Moreover, since $a + 0 = 0 + a = a$ for every $a \in \mathbf{Z}$, we see that 0 is the identity for addition; and clearly 1 is the identity for multiplication.

On the same set **Z**, let us define $a \circ b = a - b$. Since $3 \circ 2 = 1$ and $2 \circ 3 = -1$, we see that this operation is not commutative. Note that just *one* instance in which $a \circ b \neq b \circ a$ implies that the operation is not commutative. The reader may verify that neither is this operation associative. Does there exist an identity for this operation? Since $a \circ 0 = a - 0 = a$ for every integer a, it might appear at first glance that 0 is an identity. However, $0 \circ a = -a$ and the definition of an identity is not met.

EXERCISES

1. On the set of all nonempty subsets of a nonempty set A, consider the relation R defined by aRb if and only if $a \cap b \neq \emptyset$. Which of the three defining properties 1.8 of an equivalence relation hold for this relation?
2. Let R be the relation on the set $\{1, 2\}$ defined by $1R1$ and the relation R holds for no other ordered pair except the pair $(1, 1)$. Show that the relation R has exactly two of the defining properties of an equivalence relation.

Acc. no – – 16691

3. Give an example of a relation R on some set such that R has the symmetric property but does not have the reflexive property or the transitive property.

4. If "$\sim$" is an equivalence relation on a set A, carefully prove each of the following:

   (i) If $a, b \in A$ such that $a \sim b$, then $[a] \cap [b] = \varnothing$.
   (ii) If $a, b, c, d \in A$ such that $c \in [a]$, $d \in [b]$, and $[a] \neq [b]$, then $c \sim d$.

5. If a and b are integers, let us define $a \equiv b$ to mean that $a - b$ has 5 as a factor. Verify that "$\equiv$" is an equivalence relation on the set $\mathbf{Z}$ of all integers, and exhibit all the different equivalence classes.

6. In each of the following, "$\circ$" is the specified binary operation on the set $\mathbf{Z}$ of integers. Determine in each case whether the operation is commutative, whether it is associative, and whether there is an identity for the operation.

   (i) $a \circ b = b$,
   (ii) $a \circ b = a + b + ab$,
   (iii) $a \circ b$ is the larger of a and b,
   (iv) $a \circ b = 2a + 2b$,
   (v) $a \circ b = a + b - 1$,
   (vi) $a \circ b = a + ab$.

7. Let $\mathscr{P}(A)$ be the power set of a set A.

   (i) Is the binary operation "$\cap$" on $\mathscr{P}(A)$ commutative? Is it associative? Does it have an identity?
   (ii) Answer the same questions for the binary operation "$\cup$" on $\mathscr{P}(A)$.

8. Let $A = \{e, x\}$. Show that it is possible to define sixteen operations on A. Furthermore, show that e is an identity for only two of these operations.

Chapter 2

Rings

In this chapter we shall introduce the important class of algebraic systems that are called *rings*, give a large number of examples, and then establish some fundamental properties common to all rings. All the properties that are used to define a ring are suggested by simple properties of the integers, and we begin by pointing out some of these properties. The following section is therefore of a preliminary nature and is merely intended to furnish a partial motivation of the material to follow.

2.1 FORMAL PROPERTIES OF THE INTEGERS

The simplest numbers are the numbers 1, 2, 3, …, used in counting. These are called the "natural numbers" or the "positive integers." Addition and multiplication of natural numbers have simple interpretations if we consider a natural number as indicating the number of elements in a set. For example, suppose that we have two piles of stones, the first one containing m stones and the second one n stones. If the stones of the first pile are placed on the second pile, there results a pile of $n + m$ stones. If, instead, the stones of the second pile are placed on the first pile, we get a pile of $m + n$ stones. It thus seems quite obvious that

$$m + n = n + m$$

for every choice of m and n as natural numbers; that is, addition of natural numbers is commutative. This property of the natural numbers is an example of what is sometimes called a *law* or a *formal property*. Another example is the associative law of addition.

Multiplication of natural numbers may be introduced as follows. If one has m piles, each of which contains n stones, and all the stones are placed in one pile, the resulting pile will contain mn stones. It is also a familiar fact that multiplication of natural numbers is both commutative and associative. Moreover, addition and multiplication are such that the so-called *distributive law* holds:

$$m(n + k) = mn + mk,$$

where m, n, and k are arbitrary natural numbers.

Historically, the natural numbers were no doubt used for centuries before there was any consideration of their formal properties. However, in modern algebra it is precisely such formal properties that are of central interest. Some of the reasons for this changed viewpoint will become evident later on in this chapter as well as in succeeding chapters.

Of course, if m and n are natural numbers, there need not be a natural number x such that $m + x = n$. In order to be able to solve all equations of this kind, we need to have available the negative integers and zero along with the positive integers. The properties with which we shall be concerned in the next section are suggested by well-known properties of the system of *all* the integers (positive, negative, and zero). Near the end of the next chapter we shall be ready to give what may be called a characterization of the system of all integers, although, for the most part, we shall merely assume a familiarity with the simpler properties of this system. In later chapters, the other number systems of elementary algebra will be discussed in some detail. However, even before they are presented in a logical way we shall not hesitate to illustrate parts of our general theory by examples from these familiar number systems.

2.2 DEFINITION OF A RING

The concepts to be presented in this section are of fundamental importance, although a full realization of their generality will probably not become apparent until the examples of the following section are studied carefully.

We begin with a nonempty set R on which there are defined two binary operations, which we shall call "addition" and "multiplication," and for which we shall use the familiar notation. Accordingly, if $a, b \in R$, then $a + b$ and ab (or $a \cdot b$) are uniquely determined elements of the set R. By way of emphasis, we may state this fact in another way by saying that R is to be *closed* under the binary operations which we are calling addition and multiplication. We now assume the following properties or laws, in which a, b, and c are arbitrary elements, distinct or identical, of R.

P_1: $a + b = b + a$ (*commutative law of addition*).

P_2: $(a + b) + c = a + (b + c)$ (*associative law of addition*).

P_3: There exists an element 0 of R such that $a + 0 = a$ for every element a of R *(existence of a zero)*.

P_4: If $a \in R$, there exists $x \in R$ such that $a + x = 0$ *(existence of additive inverses)*.

P_5: $(ab)c = a(bc)$ *(associative law of multiplication)*.

P_6: $a(b + c) = ab + ac$, $(b + c)a = ba + ca$ *(distributive laws)*.

Under all these conditions R is said to be a *ring*. Let us repeat this definition in the following formal way.

2.1 Definition. If R is a nonempty set on which there are defined binary operations of addition and multiplication such that Properties P_1—P_6 hold, we say that R is a ***ring*** (with respect to these definitions of addition and multiplication).

Let us make a few remarks about the defining properties of a ring. First, we may emphasize that we should not think of the elements of a ring as necessarily being *numbers*. Moreover, addition and multiplication are not assumed to have any properties other than those specified. The element "0," whose existence is asserted in P_3, and which we call a *zero*, is actually an identity for the operation of addition since, by P_1, $0 + a = a + 0$ and therefore we also have $0 + a = a$. We do not assume that there is only *one* identity for addition, but later on we shall prove this to be true. Again, we should not think of 0 as being the familiar *number* zero; it is merely an identity for the operation of addition. Finally, we point out that P_4 does not assert that there is only *one* $x \in R$ such that $a + x = 0$, but this fact will also be proved eventually.

All the properties used to define a ring are familiar properties of the integers. Hence, with the usual definitions of addition and multiplication, the set of all integers is a ring. Henceforth, this ring will be denoted by **Z**. For this ring, the zero whose existence is asserted in P_3 is the familiar number zero.

Now let E be the set of all *even* integers (positive, negative, and zero). Using, of course, addition and multiplication as already defined in **Z**, we see that the sum of two elements of E is an element of E, and similarly for the product of two elements. That is, E is closed under the operations of addition and multiplication. Properties P_1, P_2, P_5, and P_6 hold in E since they hold in the larger set **Z**. Moreover, it is clear that P_3 and P_4 also hold in E, and therefore E is itself a ring.

If all elements of a ring S are contained in a ring R, it is natural to call S a *subring* of R. It is understood that addition and multiplication of elements of S are to coincide with addition and multiplication of these elements considered as elements of the larger ring R. Naturally, a set S of elements of R cannot possibly be a subring of R unless S is closed under the operations of addition and multiplication on R since, otherwise, we would not have operations *on the set S*. We see that E, as defined above, is a subring of the ring **Z**. However, the set of all odd integers cannot be a subring of **Z** since this set is not closed under addition; that is, the sum of two odd integers is not always (in fact, is never) an odd integer.

It is important to observe that the definition of a ring does not require that the operation of multiplication be commutative. However, we shall frequently want to consider this property, so let us give it a number as follows:

P_7: If $a, b \in R$, then $ab = ba$ (*commutative law of multiplication*).

A ring which has property P_7 is called a *commutative* ring. If P_7 does not hold, that is, if there exist at least two elements c and d of R such that $cd \neq dc$, then R is said to be a *noncommutative* ring.

We may also point out that in a ring there need not be an identity for the operation of multiplication. If in a ring R there exists an identity for multiplication, we shall usually call it a *unity* of R and say that R is a *ring with unity*. For convenience of reference, let us give this property a number as follows:

P_8: There exists an element e of R such that $ea = ae = a$ for every element a of R (*existence of a unity*).

We may emphasize that a ring need not have either of the properties P_7 or P_8. However, most of the rings that we shall study in detail will have both of these properties. The ring **Z** is an example of a commutative ring with unity, whereas the ring E of all even integers is a commutative ring without a unity. A few cases of noncommutative rings will occur among the examples of the next section. Naturally, they will have to be quite different from the familiar number systems.

2.3 EXAMPLES OF RINGS

In order to give an example of a ring R, it is necessary to specify the elements of R and to define the operations of addition and multiplication on R so that Properties P_1—P_6 hold. The ring **Z** of integers has been mentioned as a well-known example of a ring. Other examples are the ring of all real numbers and the ring of all complex numbers, with the usual definitions of addition and multiplication. It will be recalled that the *rational numbers* are those numbers which can be expressed in the form m/n, where m and n are integers with $n \neq 0$. With respect to the familiar definitions of addition and multiplication of rational numbers, the set of all rational numbers is also a ring. Clearly, the ring **Z** is a subring of the ring of all rational numbers; the ring of all rational numbers is a subring of the ring of all real numbers; and the ring of all real numbers is a subring of the ring of all complex numbers. All these number systems will be considered in detail in later chapters.

We proceed to give some other, less familiar, examples of rings. For the most part, we shall not write out the verifications of the Properties $P_1 - P_6$. Some of these verifications will be required in the next list of exercises. The purpose of these examples is to clarify the concept of a ring and to show that there are rings of many different kinds.

Example 1: Let S be the set of all real numbers of the form $x + y\sqrt{2}$, where $x, y \in \mathbf{Z}$, with addition and multiplication defined in the usual way. It may be verified that S is closed under these operations. Actually, S is a commutative ring with unity. Of course, it is a subring of the ring of all real numbers.

Example 2: Let T be the set of all real numbers of the form $u + v\sqrt[3]{2} + w\sqrt[3]{4}$, where u, v, and w are rational numbers. Using the usual definitions of addition and multiplication, T is a commutative ring with unity.

Example 3: Let $R = \{u, v, w, x\}$; that is, R consists of just these four elements. We define addition and multiplication in R by means of the following tables.

$(+)$	u	v	w	x
u	u	v	w	x
v	v	u	x	w
w	w	x	u	v
x	x	w	v	u

$(\cdot)$	u	v	w	x
u	u	u	u	u
v	u	v	w	x
w	u	w	w	u
x	u	x	u	x

These we read as follows. For example, we find $v + x$ by looking in the addition table at the intersection of the row which contains v as its left-hand element and the column which contains x at the top. Since w appears in this position, we have $v + x = w$. Other examples are: $w + w = u$, $x + w = v$, $vw = w$, $xx = x$. It would take too much calculation to verify the associative laws and the distributive laws, and we shall now merely state that they do hold. From the addition table, it is seen that the zero of the ring R is the element u; and from the multiplication table it follows that v is the unity. The reader may verify that this is a commutative ring. This ring R differs from previous examples in that it has only a finite number (four) of elements.

Example 4: Let C be the set of all functions which are continuous on the closed interval $0 \le x \le 1$. For two functions f and g in C, we define their sum and product by the rules

$$(f + g)(x) = f(x) + g(x),$$
$$(fg)(x) = f(x)g(x).$$

This way of defining the operations for a set of functions is called *pointwise addition* and *pointwise multiplication*. The axioms required of a ring may now be checked. For example, the function defined by $z(x) = 0$ for all x is the zero required by axiom P_3. Does the ring C have a unity?

Example 5: The set $T = \{0, e\}$ is a ring of two elements if addition and multiplication are defined by the following tables.

$(+)$	0	e
0	0	e
e	e	0

$(\cdot)$	0	e
0	0	0
e	0	e

Clearly, 0 is the zero of this ring and e is the unity. Hence, this ring has *only* a zero and a unity.

Example 6: Let $K = \{a, b, c, d\}$, with addition and multiplication defined by the following tables.

$(+)$	a	b	c	d
a	a	b	c	d
b	b	a	d	c
c	c	d	a	b
d	d	c	b	a

$(\cdot)$	a	b	c	d
a	a	a	a	a
b	a	b	c	d
c	a	a	a	a
d	a	b	c	d

The ring K is our first example of a noncommutative ring. From the multiplication table we see, for example, that $cd = a$, whereas $dc = c$. Does this ring have a unity? What is the zero?

We may emphasize that in this example, as in others in which addition and multiplication of more than two elements are defined by tables, it would be exceedingly tedious to verify the associative and distributive laws. Of course, the tables have not been written down at random but have been obtained by methods not yet available to the student. At present, the associative and distributive laws will have to be taken on faith, but there is no real difficulty in verifying the other defining properties of a ring.

Example 7: For later reference, we give still another example of a ring with four elements a, b, c, and d. In this case, we define addition and multiplication as follows:

$(+)$	a	b	c	d
a	a	b	c	d
b	b	a	d	c
c	c	d	a	b
d	d	c	b	a

$(\cdot)$	a	b	c	d
a	a	a	a	a
b	a	b	c	d
c	a	c	d	b
d	a	d	b	c

It will be observed that the addition table coincides with the addition table of the preceding example. However, the multiplication table is quite different. This ring is another example of a commutative ring.

Example 8: Let L be the set $\mathbf{Z} \times \mathbf{Z} \times \mathbf{Z}$. That is, L is the set of all ordered triples (a, b, c), where $a, b, c \in \mathbf{Z}$. We make the following definitions:

$$(a,b,c) + (d,e,f) = (a+d, b+e, c+f),$$
$$(a,b,c)(d,e,f) = (ad, bd+ce, cf).$$

To avoid any possible confusion, we may again state that we consider two elements of a set to be equal only if they are identical. Hence, if (a,b,c) and (d,e,f) are elements of L, then $(a,b,c) = (d,e,f)$ means that $a = d$, $b = e$, and $c = f$.

It is easy to verify that $(0,0,0)$ is the zero of the ring L, and that $(1,0\ 1)$ is a unity. This is another noncommutative ring since, for example,

$$(0,1,0)(1,0,0) = (0,1,0),$$

whereas

$$(1,0,0)(0,1,0) = (0,0,0).$$

Let us verify one of the distributive laws for this ring. If (a,b,c), (d,e,f), and (g,h,i) are elements of L, let us show that

$$(a,b,c)((d,e,f) + (g,h,i)) = (a,b,c)(d,e,f) + (a,b,c)(g,h,i).$$

The equality of these expressions is a consequence of the following simple calculations:

$$\begin{aligned}(a,b,c)((d,e,f) + (g,h,i)) &= (a,b,c)(d+g, e+h, f+i)\\ &= (a(d+g), b(d+g) + c(e+h), c(f+i))\end{aligned}$$

and

$$\begin{aligned}&(a,b,c)(d,e,f) + (a,b,c)(g,h,i)\\ &\qquad = (ad, bd+ce, cf) + (ag, bg+ch, ci)\\ &\qquad = (ad+ag, (bd+ce) + (bg+ch), cf+ci).\end{aligned}$$

The right sides of these equations are equal in view of certain simple properties of the integers. What properties are involved?

Example 9: Let $M_2(\mathbf{Z})$ be the set of all symbols of the form

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

where a, b, c, and d are arbitrary elements of $\mathbf{Z}$. Our definitions of addition and multiplication are as follows:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix},$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}.$$

With respect to these definitions of addition and multiplication, $M_2(\mathbf{Z})$ is a ring. It is called the *ring of all two-by-two matrices over the integers*. The reader may verify, by examples, that the commutative law of multiplication does not hold and hence that $M_2(\mathbf{Z})$ is a noncommutative ring.

We may point out that the elements of $M_2(\mathbf{Z})$ are quadruples of elements of $\mathbf{Z}$, and could just as well have been written in the form (a, b, c, d). However, the above notation is more convenient and is the traditional one.

If we modify this example by letting a, b, c, and d be rational (or real, or complex) numbers instead of integers, we obtain *the ring of all two-by-two matrices over the rational* (*or real or complex*) *numbers.*

Example 10: This final example is of a type quite different from any of the previous examples. Let A be a given set, and let R be the power set of A, that is, the set of all subsets of A, including the empty set and the entire set A. We shall now denote elements of R by lowercase letters—even though they are sets of elements of A.

Our definitions of addition and multiplication are as follows. If $a, b \in R$, we define $a + b = (a \cup b)\backslash(a \cap b)$ and $ab = a \cap b$. Note that $a + b$ consists of the elements of A that are in either subset a or in subset b, but not in both. Thus $a + b$ is not, in general, the union of a and b, but it will be this union whenever $a \cap b = \varnothing$. In the Venn diagram shown in Figure 6, region 1 represents those elements of A which are in neither subset a nor subset b, region 2 represents those elements of A which are in a but not in b, and so on. Hence ab is represented by region 4 and $a + b$ by the union of regions 2 and 3.

We now assert that with the above definitions of addition and multiplication, R is a commutative ring with unity.

The commutative laws of addition and multiplication are obvious, as is the associative law of multiplication. Let us briefly consider the associative law of addition, and let a, b, and c be arbitrary elements of R. In Figure 7, $a + b$ is represented by regions 2, 3, 5, and 6. Since c is made up of regions 4, 5, 6, and 8, it follows that $(a + b) + c$ is represented by the union of regions 2, 3, 4, and 8. This pictorial representation suggests that $(a + b) + c$ consists of those elements of A which are in exactly one of the subsets a, b, and c together with those which are in all three. To complete the verification of the associative law of addition by means of Venn diagrams, we need to characterize the set $a + (b + c)$. We omit the details, but

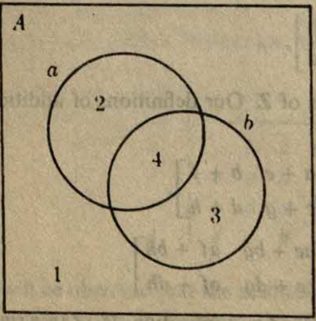


Figure 6

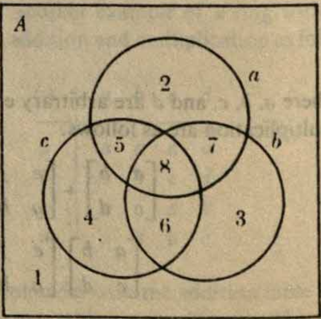


Figure 7

it is not difficult to verify that we again get the set represented by regions 2, 3, 4, and 8. Hence, $(a + b) + c = a + (b + c)$, as we wished to show. In an exercise below the reader is asked to consider how one could turn this geometrical argument into a formal proof.

If we denote the empty set by "0," it follows that $a + 0 = a$, and the empty set is the zero of the ring R. Moreover, the subset of A consisting of A itself is the unity of the ring. (Why?) If $a \in R$, it is interesting to observe that $a + a = 0$, and thus a is its own additive inverse. Another unusual property of this ring is that $a \cdot a = a$ for every element a of R. We shall refer to this ring as the *ring of all subsets of the set A.* We may emphasize that later on whenever we mention the ring of all subsets of a set *it is always to be understood that addition and multiplication are defined as in this example.*

Direct Sums

We conclude this section not by giving still another example of a ring but by presenting a simple, but quite useful, way to construct new rings from given rings. Suppose that R and S are rings, distinct or identical, and let us consider the Cartesian product $R \times S$ whose elements are the ordered pairs (r, s), $r \in R$, $s \in S$. On this set $R \times S$, we define addition and multiplication as follows:

$$(r_1, s_1) + (r_2, s_2) = (r_1 + r_2, s_1 + s_2),$$
$$(r_1, s_1)(r_2, s_2) = (r_1 r_2, s_1 s_2).$$

It is understood, of course, that $r_1, r_2 \in R$ and that $s_1, s_2 \in S$. Moreover, although the same symbol for addition is used in both rings, $r_1 + r_2$ is the sum of r_1 and r_2 in the ring R and $s_1 + s_2$ is the sum of s_1 and s_2 in the ring S (and similarly for products). We leave as an exercise the proof that with respect to the above definitions the set $R \times S$ becomes a ring. It is convenient to have a name for the ring obtained in this way. Accordingly, we make the following definition.

2.2 Definition. If R and S are given rings, the ring whose elements are the elements of the product set $R \times S$, with addition and multiplication as defined above, is called the ***direct sum*** of the rings R and S, and is usually denoted by $R \oplus S$.

What conditions on R and S will assure us that $R \oplus S$ is commutative? That it has a unity?

EXERCISES

In these exercises, it is to be assumed that the real numbers (in particular, the rational numbers and the integers) have all the familiar properties which are freely used in elementary algebra.

1. Which of the following are rings with respect to the usual definitions of addition and multiplication? In this exercise, the ring of all even integers is denoted by $\mathbf{E}$.
   (a) The set of all positive integers.
   (b) The set of all integers (positive, negative, and zero) that are divisible by 3.
   (c) The set of all real numbers of the form $x + y\sqrt{2}$, where $x, y \in \mathbf{E}$.
   (d) The set of all real numbers of the form $x + y\sqrt[3]{2}$, where $x, y \in \mathbf{Z}$.
   (e) The set of all real numbers of the form $x + y\sqrt[3]{2} + z\sqrt[3]{4}$, where $x, y, z \in \mathbf{Z}$.
   (f) The set of all real numbers of the form $x + y\sqrt{3}$, where $x \in \mathbf{E}$ and $y \in \mathbf{Z}$.
   (g) The set of all rational numbers that can be expressed in the form m/n, where $m \in \mathbf{Z}$ and n is a positive odd integer.

2. What is the additive inverse of each element of the ring R of Example 3?

3. Verify that the subset $S = \{u, w\}$ of the ring R of Example 3 is a subring of R. Show that, except for the notation employed, this is the ring of Example 5.

4. For the ring R of Example 3, use the tables to verify each of the following:

$$\begin{aligned}(u + v) + w &= u + (v + w),\\ (v + w) + x &= v + (w + x),\\ w(v + x) &= wv + wx,\\ (w + v)x &= wx + vx,\\ (xv)w &= x(vw).\end{aligned}$$

5. For the ring L of Example 8, verify the other distributive law and the associative law of multiplication.

6. For the ring $M_2(\mathbf{Z})$ of Example 9, verify the associative law of multiplication and the distributive laws. What is the zero of this ring? Verify that

$$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

is a unity of $M_2(\mathbf{Z})$. Give examples to show that $M_2(\mathbf{Z})$ is a noncommutative ring.

7. For the ring R of Example 10, consider how a formal proof of the associative law of addition could be given without use of Venn diagrams, and write out at least a part of the proof.

8. For the ring R of Example 10, use Venn diagrams to verify that if $a, b, c \in R$, then $a(b + c) = ab + ac$. How do you know without further calculation that the other distributive law must also hold?

9. On the set $S = \mathbf{Z} \times \mathbf{Z}$, let us define addition and multiplication as follows:

$$\begin{aligned}(a, b) + (c, d) &= (a + c, b + d),\\ (a, b)(c, d) &= (ac + 2bd, ad + bc).\end{aligned}$$

Prove that S is a commutative ring with unity.

10. It can be shown that the set $\{a, b, c, d\}$ is a ring if addition and multiplication are defined by the following tables.

(+)	a	b	c	d
a	a	b	c	d
b	b	c	d	a
c	c	d	a	b
d	d	a	b	c

(·)	a	b	c	d
a	a	a	a	a
b	a	c	a	c
c	a	a	a	a
d	a	c	a	c

Is this a commutative ring? Does it have a unity? What is the zero of this ring? What is the additive inverse of each element of this ring?

11. Show that neither of the following can possibly be the addition table for a ring consisting of the set $\{a,b,c,d\}$ of four elements.

(+)	a	b	c	d
a	a	b	c	a
b	b	c	d	a
c	c	d	a	b
d	a	a	b	c

(+)	a	b	c	d
a	a	b	c	d
b	b	c	a	d
c	c	a	d	b
d	b	c	d	a

12. If a ring has exactly n elements, then its addition table is a square array of n rows and n columns with the ring elements serving as indices for the rows and columns. Based upon an examination of the tables in this section, one might guess that each ring element appears exactly once in each row and in each column of the addition table. Prove that this is indeed the case for every ring.

13. Define addition of integers in the usual way, but define the "product" of any two integers to be zero. Is the set of all integers a ring with respect to addition and this new "multiplication"?

14. Let R be an arbitrary ring and consider matrices

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

where $a, b, c, d \in R$. If addition and multiplication are defined as in Example 9, prove that we obtain a ring. This ring is called the ring of all two-by-two matrices over the ring R.

15. Prove that if a ring R contains elements s and t such that $st \neq 0$, then the ring of all two-by-two matrices over R is a noncommutative ring.

16. Make addition and multiplication tables for the ring of all subsets of the set $A = \{1,2\}$. Verify that by a proper choice of notation this ring is the ring of Example 3.

17. The following is an addition table and part of the multiplication table for a ring with three elements. Make use of the distributive laws to fill in the rest of the multiplication table.

(+)	a	b	c
a	a	b	c
b	b	c	a
c	c	a	b

(·)	a	b	c
a	a	a	a
b	a	b	–
c	a	–	–

Is this a commutative ring? Does it have a unity?

18. Do the same as in the preceding exercise, using the following addition table and partial multiplication table for a ring with four elements.

(+)	a	b	c	d
a	a	b	c	d
b	b	a	d	c
c	c	d	a	b
d	d	c	b	a

(·)	a	b	c	d
a	a	a	a	a
b	a	–	–	a
c	a	–	c	–
d	a	b	c	–

Is this a commutative ring? Does it have a unity?

19. If a and b are any integers let us give the following new definitions of "addition" and "multiplication," indicated respectively by "$\oplus$" and "$\odot$":

$$a \oplus b = a + b - 1, \quad a \odot b = a + b - ab.$$

Verify that with respect to these definitions of "addition" and "multiplication" the set of all integers is a commutative ring with unity. What is the zero of this ring?

20. If R and S are rings, give a detailed proof that the direct sum $R \oplus S$ is a ring.

21. (i) If the ring R has m elements and the ring S has n elements (m and n being positive integers), how many elements are there in the direct sum $R \oplus S$?

(ii) Give an example of a commutative ring with 16 elements and an example of a noncommutative ring with 16 elements.

(iii) Give an example of a ring with 32 elements which does not have a unity.

2.4 SOME PROPERTIES OF ADDITION AND MULTIPLICATION

So far we have given the definition of a ring and have presented a number of examples of rings of many different kinds. By now it should be clear that when we think of an arbitrary ring we should not necessarily think of one of our familiar number systems. Accordingly, we cannot consider any properties of a ring as being obvious, except those actually used in the definition. In this section we shall give proofs of a number of properties of any ring. At first, we consider only properties of addition, and hence will use in our proofs only the properties $P_1 - P_4$.

Before proceeding, it may be well to recall that if a and b are elements of a set (in particular, of a ring), by $a = b$ we mean that a and b are identical elements of the set or, looked at another way, a and b are different symbols for the same element. As a consequence of this usage of "equality," it is clear that equality is an equivalence relation. Furthermore, when working in a ring, the definition of a binary operation implies that if $a = b$ and $c = d$, then $a + c = b + d$ and $ac = bd$. In the proofs below, we shall freely use these facts without explicit mention.

First, let us prove the following result.

2.3 Theorem. *The zero of a ring R, whose existence is asserted by* P_3, *is unique.*

PROOF. By this statement, we mean the following. If $0, 0' \in R$ such that for every element a of R,

$$a + 0 = a \tag{1}$$

and also

$$a + 0' = a, \tag{2}$$

then $0 = 0'$. The proof is as follows. Since Equation (1) is true for every element a of R, we may replace a in this equation by $0'$. Hence, we have that

$$0' + 0 = 0'. \tag{3}$$

In like manner, it follows from Equation (2) that

$$0 + 0' = 0. \tag{4}$$

Since, by the commutative law of addition, $0' + 0 = 0 + 0'$, it follows from Equations (3) and (4) that $0 = 0'$, and the proof is complete.

In view of this result, we are justified in speaking of *the* zero of a ring. An element which is not the zero may naturally be called a *nonzero* element.

We may observe that this is the first theorem which we have proved about an arbitrary ring; that is, the definition of a ring has this result as a logical consequence. Although the truth of this theorem may be easily verified for all the specific examples of rings which have been given, the verification for any number of examples would not in itself constitute a proof that it is always true. Now we know that it must be true in *every* ring. A considerable number of other results for arbitrary rings will be obtained as we proceed.

We next prove the following theorem.

2.4 Theorem (Cancellation Laws of Addition). *If a, b, and c are elements of a ring R, the following are true:*

(*i*) *If $a + c = b + c$, then $a = b$,*
(*ii*) *If $c + a = c + b$, then $a = b$.*

PROOF. We proceed to prove the first statement of this theorem. Let us therefore assume that

$$a + c = b + c. \tag{5}$$

By P_4, there exists an element t of R such that

$$c + t = 0. \tag{6}$$

Now it follows from Equation (5) that

$$(a + c) + t = (b + c) + t. \tag{7}$$

But

$$\begin{aligned}(a + c) + t &= a + (c + t) && (\textit{assoc. law})\\ &= a + 0 && (\textit{Equation (6)})\\ &= a && (\textit{definition of } 0).\end{aligned}$$

Similarly,

$$\begin{aligned}(b + c) + t &= b + (c + t)\\ &= b + 0 = b.\end{aligned}$$

From these calculations, and Equation (7), we see that $a = b$, as we wished to show.

In view of the commutative law of addition, part (ii) of the theorem follows at once from part (i).

Although the definition of the zero requires that $a + 0 = a$ for *every* element a of a ring R, we can now observe that the zero of a ring is completely determined by any *one* element. By this statement, we mean that *if d is some one element of R and $d + z = d$, then we must have $z = 0$.* Of course, this fact is an immediate consequence of the preceding theorem. For $d + z = d$ and $d + 0 = d$ imply that $d + z = d + 0$, from which it follows that $z = 0$.

The next result is also an almost immediate consequence of the preceding theorem.

2.5 Corollary. *The additive inverse of an element a of a ring R, whose existence is asserted by Property $\mathbf{P}_4$, is unique.*

PROOF. To prove this statement, suppose that $a + x = 0$ and that $a + y = 0$. Then $a + x = a + y$, and one of the cancellation laws of addition shows at once that $x = y$.

Since each element a of R has exactly one additive inverse, we shall find it convenient to denote this additive inverse by $-a$, and shall also often write $b + (-a)$ in the form $b - a$. It may be helpful to have in mind a verbal definition of $-a$ as follows, "$-a$ is the element of R which when added to a gives 0." That is, if $a + x = 0$ (or, equally well, $x + a = 0$), it follows that $x = -a$.

Since $a + (-a) = 0$, we see also that a is the additive inverse of $-a$, that is, that $-(-a) = a$. We have thus established the first of the following, where a, b, and c are arbitrary elements of a ring:

2.6

(i) $$-(-a) = a,$$

(ii) $$-(a + b) = -a - b,$$

(iii) $$-(a - b) = -a + b,$$

(iv) $$(a - b) - c = a - (b + c).$$

Let us prove the second of these statements. Now $-(a + b)$ is, by definition, the additive inverse of $a + b$, and we proceed to verify as follows that also $-a - b$ is the additive inverse of $a + b$:

$$\begin{aligned}
(a + b) + (-a - b) &= (a + b) + ((-a) + (-b)) && \textit{(notation)}\\
&= [(a + b) + (-a)] + (-b) && \textit{(assoc. law)}\\
&= [a + (b + (-a))] + (-b) && \textit{(assoc. law)}\\
&= [a + (-a + b)] + (-b) && \textit{(comm. law)}\\
&= [(a + (-a)) + b] + (-b) && \textit{(assoc. law)}\\
&= (0 + b) + (-b) && \textit{(def. of } -a\textit{)}\\
&= b + (-b) && \textit{(def. of 0)}\\
&= 0 && \textit{(def. of } -b\textit{)}.
\end{aligned}$$

We therefore see that both $-(a + b)$ and $-a - b$ are additive inverses of $a + b$. Hence, the uniqueness of the additive inverse implies that

$$-(a + b) = -a - b,$$

and the proof is complete. The proofs of the other two parts of 2.6 will be given as exercises in the next list of exercises.

The next theorem of this section is the following.

2.7 Theorem. *If a and b are elements of a ring R, the equation $a + x = b$ has in R the unique solution $x = b - a$.*

PROOF. It is easy to verify that $x = b - a$ is a solution. For

$$\begin{aligned} a + (b - a) &= a + (-a + b) \quad (\textit{comm. law}) \\ &= (a + (-a)) + b \quad (\textit{assoc. law}) \\ &= 0 + b = b. \end{aligned}$$

The *uniqueness* of the solution follows from one of the cancellation laws. For if we have $a + x = b$ and $a + y = b$, then $a + x = a + y$, and this implies that $x = y$.

Next we shall establish some properties of a ring that involve multiplication only and some that involve both addition and multiplication. First, we prove the following result.

2.8 Theorem. *A ring can have at most one unity.*

PROOF. The proof is much like the proof of the uniqueness of the zero. Suppose that $e, e' \in R$ such that for every elements a of R,

$$(1) \qquad ea = ae = a,$$

and also

$$(2) \qquad e'a = ae' = a.$$

In particular, Equation (1) must hold for $a = e'$, that is, we must have

$$(3) \qquad ee' = e'e = e'.$$

Similarly, by setting $a = e$ in Equation (2), we obtain

$$(4) \qquad e'e = ee' = e.$$

Equations (3) and (4) then imply that $e = e'$, and there is only one unity. If a ring has a unity, we may therefore properly speak of *the* unity of a ring.

We next make the following definition.

2.9 Definition. Let a be an element of a ring R with unity e. If there exists an element s of R such that

$$as = sa = e,$$

then s is called a ***multiplicative inverse*** of a.

One of the defining properties of a ring states that every element has an additive inverse. However, simple examples show that the situation may be quite different for multiplicative inverses. In the ring of all real numbers it is true that every nonzero element has a multiplicative inverse. In the ring Z of all integers, there are exactly two elements that have multiplicative inverses, namely, 1 and -1. In the ring of all subsets of a given set A (Example 10), the only element that has a multiplicative inverse is the unity e of the ring, that is, the subset consisting of the entire set A. For if a and b are elements of this ring, ab (which we defined to be $a \cap b$) is a proper subset of A if either a or b is a proper subset of A. Hence, $ab = e$ only if $a = e$ and $b = e$.

In view of these examples, it is clear that we must never take it for granted that an element of a ring necessarily has a multiplicative inverse. However, the following result is easy to establish.

2.10 Theorem. *If an element a of a ring R with unity e has a multiplicative inverse, it is unique.*

PROOF. Suppose that both s and t are multiplicative inverses of the element a. Then, using the fact that $sa = e$ and the associative law of multiplication, we see that

$$s(at) = (sa)t = et = t.$$

But since $at = e$, it is also true that

$$s(at) = se = s,$$

and it follows that $s = t$.

In case a has a multiplicative inverse, it is customary to designate this multiplicative inverse by a^{-1}.

It will be recalled that the zero of a ring has been defined in terms of addition only. However, we shall now prove the following theorem, which has a familiar form.

2.11 Theorem. *For each element a of a ring R, we have*

$$a \cdot 0 = 0 \cdot a = 0.$$

PROOF. Since $a + 0 = a$, it follows that

$$a(a + 0) = a \cdot a.$$

But, by one of the distributive laws,

$$a(a + 0) = a \cdot a + a \cdot 0.$$

Hence,

$$a \cdot a + a \cdot 0 = a \cdot a.$$

Now we know that $a \cdot a + 0 = a \cdot a$ and, by Theorem 2.4, we conclude that $a \cdot 0 = 0$.

In case R is a commutative ring, it follows from what we have just proved that also $0 \cdot a = 0$. If R is not commutative, a proof that $0 \cdot a = 0$ can easily be given using the other one of the distributive laws. This proof will be left as an exercise.

The following can now be verified in turn for arbitrary elements a, b, and c of a ring:

(i) $a(-b) = -(ab)$,
(ii) $(-a)b = -(ab)$,
2.12 (iii) $(-a)(-b) = ab$,
(iv) $a(b - c) = ab - (ac)$,
(v) $(b - c)a = ba - (ca)$.

The proof of (i) goes as follows. We have

$$a(b + (-b)) = a \cdot 0 = 0.$$

However, by one of the distributive laws, we know that

$$a(b + (-b)) = ab + a(-b).$$

Hence,

$$ab + a(-b) = 0.$$

But since ab has a unique additive inverse $-(ab)$, it follows that $a(-b) = -(ab)$. The proofs of the other parts of 2.12 will be left as exercises.

In view of 2.12(i) and (ii), we see that

$$-(ab) = (-a)b = a(-b).$$

Accordingly, in later sections we shall usually write simply $-ab$ for any one of these equal expressions.

Subrings

Let us now make a few remarks about the concept of *subring*. If S is a subring of R, then not only is the set S a subset of the set R but also S must be closed under the operations of addition and multiplication already defined in

R. In particular, it follows that the zero of R is also the zero of S and, moreover, the additive inverse of an element of the subring S is identical with the additive inverse of this element considered as an element of R. (Why?) Several of the defining properties of a ring hold in S simply because they hold in R. For example, the associative law of addition clearly holds in S because it holds in R and $S \subseteq R$.

The following theorem furnishes a convenient way to determine whether a set of elements of R is actually a subring of R.

2.13 Theorem. *Let R be a ring and S a nonempty subset of the set R. Then S is a subring of R if and only if the following condition holds: Whenever a, b are in S, then ab and $a - b$ are in S.*

PROOF. It is necessary to show that S is closed under addition. Let a be any element of S. The assumption made in the theorem implies that $a - a = 0$ is an element of S. Next for any b in S, we have $0 - b = -b$ is in S. Now finally for any elements a, b in S, we have $-b$ is in S and so consequently $a - (-b) = a + b$ is in S. Thus S is closed under addition, and consequently S is a subring.

EXERCISES

1. Prove 2.6(iii) and (iv).
2. If a and b are elements of the ring of all subsets of a given set (Example 10), show that (i) $a = -a$ and (ii) the equation $a + x = b$ has the solution $x = a + b$.
3. Complete the proof of Theorem 2.11 by showing that $0 \cdot a = 0$ for every element a of any ring.
4. Prove 2.12(ii)–(v).
5. If a, b, c, and d are elements of a ring, prove each of the following:

(i) $(a + b)(c + d) = (ac + ad) + (bc + bd)$,
(ii) $(a + b)(c + d) = (ac + bc) + (ad + bd)$,
(iii) $(a - b)(c - d) = (ac + bd) - (bc + ad)$,
(iv) $(a + b)(c - d) = (ac + bc) - (ad + bd)$,
(v) $(a - b)(c + d) = (ac + ad) - (bc + bd)$,
(vi) $(a(-b))(-c) = a(bc)$.

6. Verify that every nonzero element of the ring of Example 7 has a multiplicative inverse.
7. Show that an element (a, b, c) of the ring L of Example 8 has a multiplicative inverse if and only if $a = \pm 1$ and $c = \pm 1$.

8. (i) Find the multiplicative inverse of each of the following elements of the ring $M_2(\mathbf{Z})$ of Example 9:

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}, \quad \begin{bmatrix} 2 & 7 \\ 1 & 3 \end{bmatrix}, \quad \begin{bmatrix} 6 & 1 \\ 17 & 3 \end{bmatrix}.$$

(ii) Show that the element

$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$$

of the ring $M_2(\mathbf{Z})$ does not have a multiplicative inverse in $M_2(\mathbf{Z})$.

(iii) For a general element

$$x = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

of this ring, show that x has an inverse if and only if $ad - bc = \pm 1$.

9. Let R be a ring with unity. If a and b are elements of R that have multiplicative inverses, show that ab has a multiplicative inverse by verifying that $(ab)^{-1} = b^{-1}a^{-1}$.

10. Give an example of elements a and b of some ring such that a^{-1} and b^{-1} exist, but $(ab)^{-1} \neq a^{-1}b^{-1}$.

11. Suppose that a, c, and d are elements of a ring R and that a has a multiplicative inverse in R. Prove that if $ac = ad$ (or $ca = da$), then $c = d$. Show how Theorem 2.10 may be considered to be a special case of this result.

12. If R and S are rings, verify that the set of elements of the direct sum $R \oplus S$ of the form $(a, 0)$, where $\alpha \in R$ and 0 is the zero of S, is a subring of $R \oplus S$.

13. Give an example of a ring R having a subring S such that
    (i) R has a unity and S does not have a unity,
    (ii) R does not have a unity but S has a unity,
    (iii) R and S have the same unity,
    (iv) R and S both have unities, but they are different,
    (v) R is a noncommutative ring and S is commutative.

14. Show that the set of all elements of the ring $M_2(\mathbf{Z})$ of Example 9 of the form

$$\begin{bmatrix} x & 0 \\ y & z \end{bmatrix},$$

where $x, y, z \in \mathbf{Z}$, is a subring of the ring $M_2(\mathbf{Z})$.

15. If S and T are subrings of a ring R, show that $S \cap T$ is a subring of R. Give an example to show that $S \cup T$ need not be a subring.

16. (i) Give the addition table and multiplication table for a ring with exactly one element.

(ii) If a ring R has more than one element and has a unity e, show that $e \neq 0$.

17. Let R be an arbitrary ring and $a \in R$. Prove that the set $\{x \mid x \in R, ax = 0\}$ is a subring of R.

2.5 GENERAL SUMS AND PRODUCTS

The operations of addition and multiplication are *binary* operations; that is, they apply to *two* elements only. Let us now consider how we can give a meaning to sums or products of three or more elements of a ring.

If a_1, a_2, and a_3 are elements of a ring, let us define $a_1 + a_2 + a_3$ as follows:

2.14 $$a_1 + a_2 + a_3 = (a_1 + a_2) + a_3.$$

However, by the associative law of addition, it then follows that

2.15 $$a_1 + a_2 + a_3 = a_1 + (a_2 + a_3),$$

and therefore a sum of three elements is independent of the way parentheses might be introduced to indicate the manner of association of the elements.

Now that we have defined a sum of three elements of a ring, let us define a sum of four elements as follows.

2.16 $$a_1 + a_2 + a_3 + a_4 = (a_1 + a_2 + a_3) + a_4.$$

The associative law of addition then shows that

2.17 $$a_1 + a_2 + a_3 + a_4 = (a_1 + a_2) + (a_3 + a_4)$$

and also that

2.18 $$a_1 + a_2 + a_3 + a_4 = a_1 + (a_2 + a_3 + a_4).$$

These calculations verify that the sum of four elements is also independent of the way in which the elements may be associated. In general, if k is a positive integer such that

$$a_1 + a_2 + \cdots + a_k$$

has been defined, we define

2.19 $$a_1 + a_2 + \cdots + a_k + a_{k+1} = (a_1 + a_2 + \cdots + a_k) + a_{k+1}.$$

It should then appear that this gives us a definition of the sum of any number n of elements of a ring. Such a definition is called a *recursive* definition, and definitions of this kind will be considered more carefully in the next chapter.

We shall not write out the details, but in precisely the same way it is possible to give a recursive definition of a product of any number n of elements of a ring.

Exponents

We next observe that positive integral exponents may be defined in any ring R in the usual way. If a is an arbitrary element of R, we may define $a^1 = a$, $a^2 = a \cdot a$, and, in general, if k is a positive integer such that a^k has been defined, we define $a^{k+1} = a^k \cdot a$. The following familiar laws of exponents now hold, where m and n are arbitrary positive integers:

2.20 (i) $$a^m \cdot a^n = a^{m+n},$$
(ii) $$(a^m)^n = a^{mn}.$$

Suppose, now, that $a, b \in R$. Then $(ab)^2 = (ab)(ab)$, and if $ba \neq ab$, $(ab)^2$ may not be equal to a^2b^2. However, if $ba = ab$, it does follow that $(ab)^2 = a(ba)b = a(ab)b = a^2b^2$. In general, it is not difficult, using mathematical induction, to show that if R is a *commutative* ring and m is any positive integer, then

2.21 $$(ab)^m = a^m \cdot b^m.$$

We may remark that *negative* integral exponents can be defined if we restrict attention to elements which have multiplicative inverses. However, we shall postpone any consideration of negative exponents until a later chapter.

Multiples

We now introduce a convenient notation for *multiples* that parallels the exponent notation for *powers*. If $a \in R$, let us define $1a = a$, $2a = a + a$, and, in general, if k is a positive integer such that ka has been defined, we define $(k + 1)a = ka + a$. If 0 is the zero integer, we define $0a$ to be the zero element of R. Actually, there will be no confusion if the same symbol is used to designate the zero integer and the zero of the ring. Since every element of a ring has an additive inverse, we can easily introduce negative multiples as well as positive multiples. If m is a positive integer, we define $(-m)a$ to be $m(-a)$. Then $(-m)a$ is also seen to be equal to $-(ma)$. Thus, for example,

$$(-2)a = 2(-a) = -a - a = -(a + a) = -(2a).$$

The reader may easily be convinced of the truth of the following, it being understood that m and n are any integers (positive, negative, or zero) and that a and b are arbitrary elements of any ring R:

2.22 (i) $$ma + na = (m + n)a,$$
(ii) $$m(na) = (mn)a,$$
(iii) $$m(a + b) = ma + mb,$$
(iv) $$m(ab) = (ma)b = a(mb),$$
(v) $$(ma)(nb) = (mn)(ab).$$

It should perhaps be emphasized that ma is a convenient way of indicating a certain sum of elements of R. However, since the integer m is not necessarily itself an element of R, it is not correct to think of ma as the product of two elements of R. Hence, for example, 2.22(iii) is not necessarily a consequence of one of the distributive laws.

Complete proofs of many statements in this section require the use of mathematical induction. Some statements will be proved in later sections. Our present point of view will be to accept the truth of these statements. In particular, it is permissible to write sums or products of more than two elements without the use of parentheses.

EXERCISES

1. If R is a commutative ring, verify that $(ab)^3 = a^3b^3$ for all a and b in R.
2. If a is any element of the ring of Example 3, verify that $2a = 0$. (The zero of the ring is u.)
3. In the ring $M_2(\mathbf{Z})$ described in Example 9, let

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

   Verify that $(AB)^2 \neq A^2B^2$.
4. If x and y are any elements of the ring K of Example 6, verify that $(xy)^2 = x^2y^2$, even though this is not a commutative ring.
5. Let A be a nonempty set, and let R be the ring of all subsets of A as in Example 10. Show $x^2 = x$ and $2x = 0$ for every x in R.
6. A ring R is called a *Boolean ring* if $a^2 = a$ for every element a of R. If R is a Boolean ring and $a \in R$, prove that $2a = 0$. Then prove that R is necessarily a commutative ring. [Hint: Consider $(a + b)^2$.]
7. Prove that the direct sum $R \oplus S$ of two rings R and S is a Boolean ring if and only if both R and S are Boolean rings.
8. Give an example of a Boolean ring with 32 elements and an example of a ring with 32 elements which is not a Boolean ring.
9. Let R be a ring and K a subring of R with the special property that if $c \in K$ and $r \in R$, then $cr \in K$ and $rc \in K$. If $a, b \in R$, let us define $a \sim b$ to mean that $a - b \in K$. Prove each of the following:
   - (i) $\sim$ is an equivalence relation on the set R.
   - (ii) If a, b, c, and d are elements of R with $a \sim b$ and $c \sim d$, then $(a + c) \sim (b + d)$ and $ac \sim bd$.
10. If there exists an element e_1 of a ring R such that $e_1a = a$ for every element a of R, e_1 may be called a *left unity* of R. Similarly, e_2 is a *right unity* of R if $ae_2 = a$ for every element a of R. Verify that the set S of all two-by-two matrices over $\mathbf{Z}$ of the form

$$\begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix}, \qquad a, b \in \mathbf{Z},$$

is a subring of the ring $M_2(\mathbf{Z})$ of all two-by-two matrices over **Z**. Then prove each of the following:

(i) The ring S has a left unity but no right unity.
(ii) The ring S has an infinite number of left unities.

11. Using the definitions of the preceding exercise, prove that if a ring R has a *unique* left unity, it is also a right unity (and therefore the unity). [Hint: If e_1 is a left unity and $c \in R$, show that $e_1 + ce_1 - c$ also is a left unity.]

2.6 HOMOMORPHISMS AND ISOMORPHISMS

The concepts to be introduced in this section play an exceedingly important role in modern algebra. Before giving formal definitions, we illustrate the ideas by several examples.

Example 1: Let **Z** be the ring of integers and T the ring of Example 5 of Section 2.3, whose addition and multiplication tables we here reproduce for convenience.

(+)	0	e
0	0	e
e	e	0

(·)	0	e
0	0	0
e	0	e

Now let $\theta: \mathbf{Z} \to T$ be the mapping defined as follows for $i \in \mathbf{Z}$:

$$\theta(i) = \begin{cases} 0 & \text{if } i \text{ is even,} \\ e & \text{if } i \text{ is odd.} \end{cases}$$

Thus, for example, $\theta(4) = 0$ and $\theta(7) = e$. We next observe that

$$\theta(4 + 7) = \theta(11) = e = \theta(4) + \theta(7)$$

and also that

$$\theta(4 \cdot 7) = \theta(4) \cdot \theta(7).$$

Actually, it may be verified that similar results always hold. That is, if $i, j \in \mathbf{Z}$, then

$$\theta(i + j) = \theta(i) + \theta(j)$$

and

$$\theta(ij) = \theta(i)\theta(j).$$

As a matter of fact, the addition and multiplication tables for T are just those which we would get if we thought of 0 as standing for "even" and e for "odd." Thus, for example, $e + e = 0$ and "odd" + "odd" = "even."

What we have observed is that in this example "the image of a sum is the sum of the images." More precisely, if we take the sum of i and j in the ring

Z, the image of this sum is the sum in the ring T of the respective images of i and j. This fact is often expressed by saying that the operation of *addition is preserved under the mapping* θ. Similarly, the operation of multiplication is also preserved under this mapping θ.

Now the mapping θ of this example is clearly a mapping of **Z** *onto* T, and we have indicated that both the operations of addition and multiplication are preserved under this mapping. According to the definition below, the mapping θ is an example of a *homomorphism* of the ring **Z** onto the ring T.

Example 2: Let R and S be arbitrary rings, and let $\phi: R \times S \rightarrow R$ be the *projection* of the set $R \times S$ onto R, as defined in Example 2 of Section 1.2. Since the set $R \times S$ becomes a ring $R \oplus S$ under natural definitions of addition and multiplication, ϕ may be considered as a mapping of the ring $R \oplus S$ onto the ring R defined by

$$\phi(r,s) = r, \qquad (r,s) \in R \oplus S.$$

We assert that the operations of addition and multiplication are preserved under the mapping ϕ. That this is true for addition is a consequence of the following simple calculations:

$$\begin{aligned}\phi[(r_1,s_1) + (r_2,s_2)] &= \phi(r_1 + r_2, s_1 + s_2) = r_1 + r_2 \\ &= \phi(r_1,s_1) + \phi(r_2,s_2).\end{aligned}$$

A similar calculation will verify that multiplication also is preserved under the mapping ϕ. Hence ϕ is a homomorphism of $R \oplus S$ onto R.

Example 3: Let $K = \{a,b,c,d\}$ be the ring of Example 6 of Section 2.5 with addition and multiplication tables which we here reproduce.

(+)	*a*	*b*	*c*	*d*
a	*a*	*b*	*c*	*d*
b	*b*	*a*	*d*	*c*
c	*c*	*d*	*a*	*b*
d	*d*	*c*	*b*	*a*

(·)	*a*	*b*	*c*	*d*
a	*a*	*a*	*a*	*a*
b	*a*	*b*	*c*	*d*
c	*a*	*a*	*a*	*a*
d	*a*	*b*	*c*	*d*

Now let $L = \{i,j,k,l\}$, with addition and multiplication on L defined by the following tables.

(+)	*i*	*j*	*k*	*l*
i	*k*	*l*	*i*	*j*
j	*l*	*k*	*j*	*i*
k	*i*	*j*	*k*	*l*
l	*j*	*i*	*l*	*k*

(·)	*i*	*j*	*k*	*l*
i	*i*	*j*	*k*	*l*
j	*k*	*k*	*k*	*k*
k	*k*	*k*	*k*	*k*
l	*i*	*j*	*k*	*l*

It can be proved that L is a ring with respect to these definitions of addition and multiplication. At first glance, the rings K and L may seem quite different, but it is not difficult to verify that they are identical except for the notation used. If in the tables for K we replace *a* by *k*, *b* by *i*, *c* by *j*, and *d* by *l*, the tables will coincide except for the order in which the elements are written down. Let us state this fact more precisely as follows. Let $\psi: K \to L$ be the mapping defined by

$$\psi(a) = k, \quad \psi(b) = i, \quad \psi(c) = j, \quad \psi(d) = l.$$

Then not only is ψ a *one-one* mapping of K onto L, but also the operations of addition and multiplication are preserved under the mapping ψ. Thus ψ is a homomorphism of K onto L which is a one-one mapping (in contrast to the homomorphisms of the two preceding examples). Such a homomorphism is given the special name of an *isomorphism.*

Let us now give formal definitions of the concepts which have been introduced in the above examples.

2.23 Definition. A mapping $\theta: R \to S$ of a ring R into a ring S is called a ***homomorphism*** if and only if for arbitrary elements a, b, of R, the following hold:

2.24 $$\theta(a + b) = \theta(a) + \theta(b), \qquad \theta(ab) = \theta(a)\theta(b).$$

If there exists a homomorphism of R *onto* S, we may say that R is ***homomorphic*** to S or that S is a ***homomorphic image*** of R.

A brief way of referring to 2.24 is to say θ preserves addition and multiplication.

The following special case is of such importance that we give it a separate definition.

2.25 Definition. A homomorphism which is a one-one and onto mapping is called an ***isomorphism.*** If there is an isomorphism of R *onto* S, we may say that R is ***isomorphic*** to S or that S is an ***isomorphic image*** of R.

We may emphasize that, for the sake of generality, we have not required a homomorphism to be an onto mapping. However, whenever we say that R is homomorphic to S or that S is a homomorphic image of R, we mean to imply that it is an onto mapping. Although the more general concept is useful in certain parts of the theory, we shall seldom have occasion to refer to homomorphisms that are not onto mappings. The reader should watch carefully to see whether the word "onto" is used.*

---

* A homomorphism that is a one-one mapping is also called a *monomorphism.* Likewise, a homomorphism that is an onto mapping is sometimes called an *epimorphism.* Thus, to say that R is isomorphic to S is to say that there exists a mapping $R \to S$ which is both a monomorphism and an epimorphism.

It may be worth observing that there always exists a trivial homomorphism of any ring R into any ring S. We have only to define the image of every element of R to be the zero element of S. Of course, this does not assert that S is a homomorphic image of R.

If the mapping $\theta: R \to S$ is an isomorphism of the ring R onto the ring S, it may be verified that the mapping $\theta^{-1}: S \to R$, as defined in Section 1.2, is an isomorphism of S onto R. Accordingly, if R is isomorphic to S, then S is isomorphic to R; we may sometimes simply say that R and S are *isomorphic rings*. As suggested by the last example above, it should be clear that isomorphic rings may be considered as differing only in the notation used to indicate the elements of the rings. Accordingly, isomorphic rings are sometimes said to be *abstractly identical*.

The most fundamental properties of homomorphisms are stated in the following theorem. It should be kept in mind that an isomorphism is a special case of a homomorphism, so that isomorphisms certainly have the stated properties.

2.26 Theorem. *Let $\theta: R \to S$ be a homorphism of the ring R into the ring S. Then each of the following is true:*

(*i*) *If 0 is the zero of R, then $\theta(0)$ is the zero of S.*
(*ii*) *If $a \in R$, then $\theta(-a) = -\theta(a)$.*
(*iii*) *If R has a unity e and θ is an onto mapping, then S has $\theta(e)$ as unity.*
(*iv*) *Suppose that R has a unity and that θ is an onto mapping. If a is an element of R having a multiplicative inverse, then $\theta(a^{-1}) = \theta(a)^{-1}$.*
(*v*) *If R is a commutative ring and θ is an onto mapping, then S is a commutative ring.*

PROOF OF (iii). We need to show that $s\theta(e) = \theta(e)s = s$ for every element s of S. Let s be an arbitrary element of S. Since θ is an onto mapping, there exists at least one element r of R such that $\theta(r) = s$. Now e is a unity of R, and therefore $re = er = r$. Hence $\theta(re) = \theta(er) = \theta r$. Since multiplication is preserved under the mapping θ, it follows that $\theta(r)\theta(e) = \theta(e)\theta(r) = \theta(r)$, or $s\theta(e) = \theta(e)s = s$, as required.

We leave the proof of the other parts of the theorem as an exercise.

EXERCISES

In these exercises the examples referred to are those of Section 2.3.

1. Prove Theorem 2.26(i), (ii), (iv), and (v).
2. In the ring K of Example 6, show that $\{a, b\}$ is a subring of K which is isomorphic to the ring of Example 5.

3. If $\theta: R \to S$ is a homomorphism of the ring R into the ring S, prove that the set T of all images of elements of R is a subring of S.

4. Let **R** denote the field of real numbers and let **C** denote the ring of all real-valued functions on the interval $[0, 1]$ (Example 4). Let p be any fixed number on the interval $[0, 1]$. Define a function $\phi: \mathbf{C} \to \mathbf{R}$ by the rule $\phi(f) = f(p)$ for each function f in **C**. Show that ϕ is a homomorphism. Determine if this homomorphism is either one-to-one or onto.

5. If R and S are rings, verify that the subring of $R \oplus S$ consisting of all elements of the form $(r, 0)$, $r \in R$, is isomorphic to R.

6. Show that the set of all elements of the ring L of Example 8 of the form $(x, 0, x)$, $x \in \mathbf{Z}$, is a subring of L which is isomorphic to **Z**.

7. If L is the ring of Example 8, show that the mapping $\theta: L \to \mathbf{Z}$ defined by $\theta(a, b, c) = a$ is a homomorphism of L onto **Z**.

8. It was shown in a previous exercise that the set of all elements of the ring $M_2(\mathbf{Z})$ of Example 9 of the form

$$\begin{bmatrix} x & 0 \\ y & z \end{bmatrix}, \qquad x, y, z \in \mathbf{Z},$$

is a subring U of $M_2(\mathbf{Z})$. Show that the mapping $\theta: U \to L$ defined by

$$\theta\begin{bmatrix} x & 0 \\ y & z \end{bmatrix} = (x, y, z), \qquad x, y, z \in \mathbf{Z},$$

is an isomorphism of U onto the ring L of Example 8.

9. Let U be the ring in the previous exercise. Define $\theta: U \to \mathbf{Z}$ by

$$\theta\begin{bmatrix} x & 0 \\ y & z \end{bmatrix} = z.$$

Show that θ is a homomorphism of U onto **Z**. Is it one-to-one?

10. If R is the ring of all subsets of the set $\{1, 2\}$, and S is the ring of all subsets of the set $\{x\}$, exhibit a homomorphism of R onto S.

11. Let $X = \{1, 2, 3\}$ and $Y = \{a, b\}$. Select any function α which maps X onto Y. Let R be the ring of all subsets of X and S the ring of all subsets of Y (Example 10). Define a function from R to S by the rule

$$\theta(A) = \{\alpha(a): a \in A\}.$$

Thus θ maps a subset A of X into the subset of Y consisting of all the images under α of the elements of A. Show that θ is a homomorphism of R onto S.

12. If R and S are rings, prove that the ring $R \oplus S$ is isomorphic to the ring $S \oplus R$.

13. If $\theta: R_1 \to R_2$ and $\phi: S_1 \to S_2$ are, respectively, homomorphisms of R_1 onto R_2 and of S_1 onto S_2, exhibit a homomorphism of $R_1 \oplus S_1$ onto $R_2 \oplus S_2$.

14. Give examples of a ring R without unity and a ring S with unity such that S is a homomorphic image of R.

15. Give examples of a noncommutative ring R and a commutative ring S such that S is a homomorphic image of R.

16. Let R be a ring without unity. On the set $R \times \mathbf{Z}$ let us define addition and multiplication as follows:

$$(a, i) + (b, j) = (a + b, i + j),$$
$$(a, i)(b, j) = (ab + ja + ib, ij).$$

Prove that with respect to this addition and multiplication $R \times \mathbf{Z}$ is a ring with unity and that this ring contains a subring which is isomorphic to R.

17. Let R be the ring of Exercise 19 of Section 2.3. Then as *sets* R and $\mathbf{Z}$ are identical, but they are distinct as *rings* since the operations are different. Show that the mapping $\theta: \mathbf{Z} \to R$ defined by $\theta(a) = 1 - a$, $a \in \mathbf{Z}$, is an isomorphism of $\mathbf{Z}$ onto R.

18. If $\theta: R \to S$ is a homomorphism of the ring R onto the ring S, prove each of the following:

(i) If A is a subring of R, then the set $\theta(A) = \{\theta(a) \mid a \in A\}$ is a subring of S.
(ii) If B is a subring of S, the set C of all elements c of R such that $\theta(c) \in B$ is a subring of R.

Chapter 3

Integral Domains and Fields

The properties that were used to define a ring were suggested by simple properties of the integers. However, since we have had numerous examples of commutative rings with unity that bear little resemblance to the ring of integers, it must be the case that the ring of integers has other properties in addition to those listed in the axioms defining a commutative ring with unity. We shall give some of these additional properties in this chapter, and in some sense, complete the list of properties. By this we mean that any ring that possesses all of the listed properties will be isomorphic to the ring of integers.

One of the properties we shall require leads naturally to the method of proof by mathematical induction. Accordingly, we shall introduce this important method and use it to prove some results discussed in Section 2.5.

Our approach is not limited to the discussion of the integers; we shall introduce two classes of commutative rings—integral domains and fields—of which the integers and rational numbers are respectively examples. These two classes of rings are closely related as will be seen in the later sections.

3.1 DEFINITION OF INTEGRAL DOMAIN

We have proved that if 0 is the zero of a ring R, then $a \cdot 0 = 0 \cdot a = 0$ for every element a of R. Of course, this is a familiar property of our elementary number systems. However, in some of the rings previously mentioned there exist elements c and d, both of which are different from zero, such that $cd = 0$. For example, in the ring of Example 3 of Section 2.3 we have $wx = u$, where u is the zero. As another example, consider the ring of all subsets of a given set

(Example 10). The empty set is the zero of this ring and, by the definition of multiplication in this ring, if c and d are subsets whose intersection is the empty set, then $cd = 0$. In discussing elements of the type just mentioned, it will be convenient to make the following definition.

3.1 Definition. An element a of a ring R is said to be a ***divisor of zero in R*** if there exists a *nonzero* element c of R such that $ac = 0$ or a *nonzero* element d of R such that $da = 0$.

It is trivial that the zero of a ring R is a divisor of zero (provided R has more than one element and therefore has a nonzero element to play the role of c or d in the above definition). The possible existence of zero divisors in a ring forces one to exercise care when dealing with operations that are very familiar in the usual number systems. The cancellation laws which are frequently used when solving equations involving real numbers, for instance, do not hold in general rings. The following simple result shows the connection between zero divisors and the cancellation laws.

3.2 Theorem (CANCELLATION LAWS OF MULTIPLICATION). *If a is not a divisor of zero in a ring R, then each of the following holds:*

(*i*) *If $b, c \in R$ such that $ab = ac$, then $b = c$.*
(*ii*) *If $b, c \in R$ such that $ba = ca$, then $b = c$.*

PROOF. Let us prove (i) of this theorem. If $ab = ac$, it follows that $a(b - c) = 0$. Then, since a is not a divisor of zero, we must have $b - c = 0$ or $b = c$. Of course, part (ii) follows by a similar argument.

It is important to keep in mind that the cancellation laws of multiplication hold *only if a is not a divisor of zero.*

In most of this chapter we shall be studying rings without nonzero divisors of zero. In such a ring the cancellation laws of multiplication as stated in Theorem 3.2 always hold provided only that $a \neq 0$. Moreover, in order to restrict ourselves for the present to rings more like the ring of integers, we shall also require our rings to be commutative and to have a unity. The next definition gives a convenient way to refer to rings having all of these properties.

3.3 Definition. A ring D with more than one element is called an ***integral domain*** if it is commutative, has a unity, and has no nonzero divisors of zero.

An alternate way of stating that a ring D has no nonzero divisors of zero is to say that it has the following property:

If $r, s \in D$ such that $rs = 0$, then $r = 0$ or $s = 0$.

In view of the definition of an integral domain, the cancellation laws of multiplication (as stated in Theorem 3.2) are always valid in an integral domain so long as $a \neq 0$.

The most familiar examples of integral domains are the ring of integers, the ring of real numbers, and the ring of complex numbers. The reader may verify that the rings of Examples 1, 2, 5, and 7 of Section 2.3 are integral domains, whereas the rings of Examples 3, 4, 6, and 10 are not integral domains.

3.2 ORDERED INTEGRAL DOMAINS

One important property of the integers that has not been mentioned so far is that they can be *ordered*. If we think of the integers as being exhibited in the following way,

$$\ldots, -4, -3, -2, -1, 0, 1, 2, 3, 4, \ldots,$$

and a and b are integers, we say that "a is greater than b" if a occurs to the right of b in the above scheme. It is clear that "a is greater than b" means merely that $a - b$ is a positive integer. This observation suggests that the concept of "order" can be defined in terms of the concept of "positive." We therefore make the following definition.

3.4 Definition. An integral domain D is said to be an ***ordered integral domain*** if D contains a subset D^+ with the following properties:

(i) If $a, b \in D^+$, then $a + b \in D^+$ (*closed under addition*).

(ii) If $a, b \in D^+$, then $ab \in D^+$ (*closed under multiplication*).

(iii) For each element a of D exactly *one* of the following holds:

$$a = 0, \quad a \in D^+, \quad -a \in D^+ \quad (\textit{trichotomy law}).$$

The elements of D^+ are called the *positive* elements of D. The nonzero elements of D that are not in D^+ are called the *negative* elements of D.

We may emphasize that D^+ is just the notation used to designate a particular subset of an ordered integral domain D. No significance is to be attached to the use of the symbol "+" in this connection.

Obviously, the set $\mathbf{Z}^+$ of positive integers has the properties required of D^+ in the above definition, and hence $\mathbf{Z}$ is an ordered integral domain. However, there are other ordered integral domains such as, for example, the integral domain of all rational numbers or the integral domain of all real numbers. However, not all integral domains are ordered integral domains. For example, we shall prove later on that the integral domain of all complex numbers has no subset with the three properties listed in the preceding definition, and therefore this integral domain is not ordered. See also Exercise 20 at the end of this section.

Now let D be any ordered integral domain, and let D^+ be the set of positive elements of D, that is, the set having the three properties stated in the preceding definition. We now show how D^+ is used to define an order relation on D. If $c, d \in D$, we *define* $c > d$ (or $d < c$) to mean that $c - d \in D^+$. Then it is clear that $a > 0$ means that $a \in D^+$, that is, that a is a positive element of D. Similarly, $a < 0$ means that $-a \in D^+$ or that a is a negative element of D. The three properties of Definition 3.4 can then be restated in the following form:

3.5
(i) If $a > 0$ and $b > 0$, then $a + b > 0$.
(ii) If $a > 0$ and $b > 0$, then $ab > 0$.
(iii) If $a \in D$, then exactly one of the following holds:

$$a = 0, \quad a > 0, \quad a < 0.$$

It is now not difficult to verify the following additional properties of inequalities:

3.6
(i) If $a > b$, then $a + c > b + c$ for every $c \in D$.
(ii) If $a > b$ and $c > 0$, then $ac > bc$.
(iii) If $a > b$ and $c < 0$, then $ac < bc$.
(iv) If $a > b$ and $b > c$, then $a > c$.
(v) If $a \neq 0$, then $a^2 > 0$.

The proof of the first of these is as follows. If $a > b$, we have $a - b > 0$. However, $a + c - (b + c) = a - b$ and we see at once that $a + c - (b + c) > 0$, that is, that $a + c > b + c$.

Let us now prove 3.6(v). If $a \neq 0$, then by the form 3.5(iii) of the trichotomy law, either $a > 0$ or $-a > 0$. If $a > 0$, it follows from 3.5(ii) that $a^2 > 0$. If $-a > 0$, the same argument shows that $(-a)^2 > 0$. Since, by 2.12(iii), $(-a)^2 = a^2$, it follows again that $a^2 > 0$.

Proofs of the other parts of 3.6 will be left as exercises.

It is convenient at times to use some slight variations of the notation just introduced. The symbols $a < b$ and $b > a$ should be understood to mean the same thing. Also we can define $a \geq b$ (or $b \leq a$) to mean that either $a = b$ or $a > b$, without specifying which. We shall henceforth use this notation whenever convenient to do so. If $a \geq 0$, it is sometimes convenient to say that a is ***nonnegative***. By writing $a < b < c$, we shall mean that $a < b$ and that also $b < c$.

If e is the unity of an ordered integral domain D, we have that $e^2 = e$, $e \neq 0$, and therefore $e^2 > 0$ by 3.6(v). Thus, *in any ordered integral domain D, the unity is a positive element* (that is, it is an element of D^+).

In any ordered integral domain it is possible to introduce the concept of absolute value in the usual way as follows.

3.7 Definition. Let D be any ordered integral domain and $a \in D$. The ***absolute value*** of a, written as $|a|$, is defined as follows:

(i) If $a \geq 0$, then $|a| = a$.
(ii) If $a < 0$, then $|a| = -a$.

From this definition it follows that $|0| = 0$ and that if $a \neq 0$, then $|a| > 0$.

A few properties of the absolute value which follow from the definition are given in the exercises.

EXERCISES

1. If a is a divisor of zero in a commutative ring R, show that ar also is a divisor of zero for every element r of R.
2. Let N be the set of elements of an arbitrary ring R which are *not* divisors of zero. Prove that N is closed under multiplication, and verify by an example that N need not be closed under addition.
3. Prove that if a has a multiplicative inverse a^{-1} in a ring R, then a is not a divisor of zero in R.
4. Consider the elements of the ring $M_2(Z)$

$$\begin{bmatrix} 0 & 1 \\ 0 & 3 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, x = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, y = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Show that the first two elements are zero divisors. Show that the elements x and y are zero divisors if and only if $ad - bc = 0$. (Compare this with Exercise 8 of Section 2.4.)

5. Determine all divisors of zero in the ring $\mathbf{Z} \oplus \mathbf{Z}$.
6. If R and S are integral domains, prove that their direct sum $R \oplus S$ cannot be an integral domain.
7. Let R be the ring of all subsets of the set $\{x, y, z\}$, and $\mathbf{Z}$ the ring of integers.
   (i) Determine all those elements of $\mathbf{Z} \oplus R$ which have multiplicative inverses.
   (ii) Determine all those elements of $\mathbf{Z} \oplus R$ which are divisors of zero.
8. Let S be a subring of the ring R. If S has a unity $e \neq 0$ but R does not have a unity, prove that e is a divisor of zero in R.
9. Give at least two examples of rings R and S with the property that there exists a homomorphism θ of R onto S and an element a of R which is a divisor of zero in R but $\theta(a)$ is not a divisor of zero in S.

10. If a commutative ring R has a subset R^+ with the properties specified for D^+ in Definition 3.4, we may naturally call R an *ordered ring*.

    (i) Give an example of an ordered ring which is not an integral domain.
    (ii) Prove that if R has a nonzero divisor of zero, R cannot be an ordered ring.

11. Prove 3.6(ii), (iii), (iv).

In Exercises 12–18, the letters a, b, c, and d represent elements of an ordered integral domain.

12. Prove that if $a > b$, then $-a < -b$.

13. Prove that if $a > b$ and $c > d$, then $a + c > b + d$.

14. Prove that if a, b, c, and d are all positive with $a > b$ and $c > d$, then $ac > bd$.

15. Prove that if $a > 0$ and $ab > ac$, then $b > c$.

16. Prove that $|ab| = |a| \cdot |b|$.

17. Prove that $-|a| \leq a \leq |a|$.

18. Prove that $|a + b| \leq |a| + |b|$.

19. Prove: There cannot be a greatest element in an ordered integral domain D (that is, for each $d \in D$ there exists $c \in D$ such that $c > d$).

20. Use the result of the preceding exercise to give a convincing argument (a formal proof is not required) why an integral domain with a finite number of elements cannot be an ordered integral domain.

21. If Property 3.2(i) holds for every nonzero element a of a ring R, prove that R has no nonzero divisor of zero.

22. If Property 3.2(i) holds for every nonzero element a of a ring R, prove that Property 3.2(ii) also holds for every nonzero element a of R.

23. Prove that in a Boolean ring, as defined in Exercise 6 of Section 2.5, every nonzero element except the unity (if it has a unity) is a divisor of zero.

24. Prove that an isomorphic image of an integral domain is an integral domain.

3.3 WELL-ORDERING AND MATHEMATICAL INDUCTION

We need one further condition to characterize the ring of integers among the ordered integral domains. Let us first make the following general definition.

3.8 Definition. A set S of elements of an ordered integral domain is said to be ***well-ordered*** if each nonempty subset U of S contains a least element; that is, if for each nonempty subset U of S there exists an element a of U such that $a \leq x$ for every element x of U.

It is apparent that the set of all positive integers is well-ordered, and we shall presently find that this property is precisely what distinguishes the ring of integers from other ordered integral domains. The rational numbers will be considered in detail later on in this book, but we may observe now that the set of positive rationals is not well-ordered. In fact, the set of all positive rational numbers has no least element. For if r is any positive rational number, then $r/2$ is also a positive rational number and $r/2 < r$. Hence there can be no least positive rational number.

Let us pause to clarify our point of view about the ring of integers. We have not *proved* any property of the integers; instead, we have from time to time merely assumed that they have certain properties. We are now able to state precisely as follows just what properties of the integers we do wish to consider as known. *We assume that the ring of integers is an ordered integral domain in which the set of positive elements is well-ordered.* Accordingly, when we shall henceforth speak of a proof of any property of the integers we shall mean a proof based on this assumption only. The theorem of the next section will indicate why no other properties are required.

We pointed out in Section 3.2 that if e is the unity of an ordered integral domain D, then $e \in D^+$. If D^+ is well-ordered, we now assert that *e is the least element of D^+*. To see this, suppose that c is the least element of D^+ and that $0 < c < e$. It follows by 3.6(ii) that $0 < c^2 < c$, since $ce = c$. Thus we would have $c^2 \in D^+$, with $c^2 < c$. But this violates the assumption that c is the least element of D^+, and we have the desired conclusion. In particular, our assumption about the ring of integers assures us that 1 is the least positive integer.

The following theorem, which is the basis of proofs by mathematical induction, is just as "obvious" as the fact that the set of positive integers is well-ordered. However, in accordance with our chosen point of view, we shall give a proof of this result.

3.9 Theorem. *Let K be a set of positive integers with the following two properties:*

(i) $1 \in K$.

(ii) If k is any positive integer such that $k \in K$, then also $k + 1 \in K$.

Then K consists of the set of all *positive integers.*

PROOF. To prove this theorem, let us assume that there is a positive integer not in K, and obtain a contradiction. Let U be the set of all positive integers not in K and therefore, by our assumption, U is not empty. Then, by the well-ordering property, U must contain a least element m. Since, by (i), we have $1 \in K$, clearly $m \neq 1$ and it follows that $m > 1$ and therefore $m - 1 > 0$. Moreover, $m - 1 \in K$ since m was chosen to be the least element of U. Now, by (ii) with $k = m - 1$, we see that $m \in K$. But $m \in U$, and we have obtained the desired contradiction. The proof is therefore complete.

The most frequent application of Theorem 3.9 is to a proof of the following kind. Suppose that there is associated with each positive integer n a *statement* (or proposition) S_n, which is either true or false, and suppose we wish to prove that the statement S_n is true for every positive integer n. Let K be the set of all positive integers n such that S_n is a true statement. If we can show that $1 \in K$, and that whenever $k \in K$ then also $k + 1 \in K$, it will follow from Theorem 3.9 that K is the set of all positive integers. Since $n \in K$ means that S_n is true, we may reformulate these remarks in the following convenient form.

3.10 Induction Principle. *Suppose that there is associated with each positive integer n a statement S_n such that the following hold:*

(*i*) *S_1 is true.*
(*ii*) *If k is any positive integer such that S_k is true, then also S_{k+1} is true.*

Then S_n is true for every positive integer n.

A proof making use of the Induction Principle (or of Theorem 3.9) is usually called a proof by induction or a proof by mathematical induction.

We have proved 3.9 (and 3.10) on the assumption that the set of positive integers is well-ordered. Exercise 10 below will show that, in fact, the Induction Principle is equivalent to the requirement that the set of positive integers be well-ordered.

We may remark that there is another useful form of the Induction Principle in which condition (ii) is replaced by a somewhat different condition. (See Exercise 9 at the end of this section.)

As a first illustration of the language and notation just introduced, we consider a simple example from elementary algebra. If n is a positive integer, let S_n be the statement that

$$2 + 4 + 6 + \cdots + 2n = n(n + 1),$$

it being understood that the left side is the sum of the first n positive even integers. We now prove that S_n is true for every positive integer n, by verifying (i) and (ii) of 3.10. Clearly, S_1 is true since S_1 merely states that $2 = 1 \cdot 2$. Suppose, now, that k is any positive integer such that S_k is true, that is, such that the following is true:

$$2 + 4 + 6 + \cdots + 2k = k(k + 1).$$

Then, by adding the next even integer, $2(k + 1)$, to both sides we obtain

$$\begin{aligned} 2 + 4 + 6 + \cdots + 2k + 2(k + 1) &= k(k + 1) + 2(k + 1) \\ &= (k + 1)(k + 2). \end{aligned}$$

However, this calculation shows that S_{k+1} is true, and hence we have verified both (i) and (ii) of 3.10. The Induction Principle then assures us that S_n is true for every positive integer n.

We now consider again part of the material of Section 2.6, and we first illustrate by a simple explanation how a recursive definition really involves the Induction Principle. The recursive definition of a^n, which was given earlier, may be stated in the following formal way.

3.11 Definition. If a is an element of a ring R, we define $a^1 = a$. Moreover, if k is a positive integer such that a^k is defined, we define $a^{k+1} = a^k \cdot a$.

Now let S_n be the statement, "a^n is defined by 3.11." The Induction Principle then shows that S_n is true for every positive integer n, that is, that a^n is defined by 3.11 for every positive integer n.

Let us now prove (2.20(i)) that if m and n are arbitrary positive integers, then

3.12
$$a^m \cdot a^n = a^{m+n}.$$

Let S_n be the statement that for the positive integer n, 3.12 is true for *every* positive integer m. Then, by definition of a^{m+1}, we see that $a^m \cdot a^1 = a^{m+1}$, and hence S_1 is true. Let us now assume that k is a positive integer such that S_k is true, that is, such that

3.13
$$a^m \cdot a^k = a^{m+k}$$

for every positive integer m. Then

$$\begin{aligned} a^m \cdot a^{k+1} &= a^m \cdot a^k \cdot a && \textit{(by def. of } a^{k+1}\textit{)} \\ &= a^{m+k} \cdot a && \textit{(by 3.13)} \\ &= a^{m+k+1} && \textit{(by def. of } a^{(m+k)+1}\textit{)}. \end{aligned}$$

We have now shown that S_{k+1} is true, and the Induction Principle then assures us that S_n is true for every positive integer n. Thus we have given a formal proof of the very familiar law of exponents stated in 3.12. In the above proof we have tacitly made use of the associative law of multiplication. As a matter of fact, it would be possible to treat the equation 3.12 as a special case of a generalized associative law of multiplication.

As a further illustration of the use of mathematical induction in proving the results stated in Section 2.5, we shall prove the generalized associative law of addition. For convenience of reference, let us first restate in a slightly different notation the recursive definition of a sum of more than two elements of a ring. If l is a positive integer and $b_1, b_2, \ldots, b_{l+1}$ are elements of a ring such that

$$b_1 + b_2 + \cdots + b_l$$

is defined, we define

3.14 $$b_1 + b_2 + \cdots + b_{l+1} = (b_1 + b_2 + \cdots + b_l) + b_{l+1}.$$

Now let S_n be the statement that for arbitrary elements $a_1, a_2, \ldots, a_n$ of a ring and for each positive integer r such that $1 \leq r < n$, we have

3.15 $$(a_1 + \cdots + a_r) + (a_{r+1} + \cdots + a_n) = a_1 + a_2 + \cdots + a_n.$$

To establish the generalized associative law of addition, we need to prove that S_n is true for every positive integer n. Clearly, S_1 and S_2 are true, and we verified S_3 and S_4 in Section 2.6. We complete the proof by showing that if k is a positive integer such that S_k is true, then also S_{k+1} is true. Otherwise expressed, if S_k is true and r is an integer such that $1 \leq r < k + 1$, we shall show that

3.16 $$(a_1 + \cdots + a_r) + (a_{r+1} + \cdots + a_{k+1}) = a_1 + a_2 + \cdots + a_{k+1}.$$

The case in which $r = k$ is true at once by definition (3.14) of the right side of 3.16. Suppose, then, that $r < k$. As a special case of 3.14, we have

$$a_{r+1} + \cdots + a_{k+1} = (a_{r+1} + \cdots + a_k) + a_{k+1}.$$

This is used in the first step of the following calculation:

$$\begin{aligned}
&(a_1 + \cdots + a_r) + (a_{r+1} + \cdots + a_{k+1}) \\
&\quad = (a_1 + \cdots + a_r) + ((a_{r+1} + \cdots + a_k) + a_{k+1}) \\
&\quad = ((a_1 + \cdots + a_r) + (a_{r+1} + \cdots + a_k)) + a_{k+1} && \textit{(by assoc. law)} \\
&\quad = (a_1 + \cdots + a_k) + a_{k+1} && \textit{(by } S_k\textit{)} \\
&\quad = a_1 + \cdots + a_{k+1} && \textit{(by 3.14)}.
\end{aligned}$$

This calculation establishes 3.16 and completes the proof.

In a similar manner the other results that were stated in Section 2.5 can be established by induction. Some of them are listed in the following set of exercises.

EXERCISES

1. Prove the generalized distributive law:

$$b(a_1 + a_2 + \cdots + a_n) = ba_1 + ba_2 + \cdots + ba_n.$$

2. Prove (2.20(ii)) that for arbitrary positive integers m and n,

$$(a^m)^n = a^{mn}.$$

3. If a and b are elements of a commutative ring, prove (2.21) that $(ab)^m = a^m b^m$ for every positive integer m.

4. If n is a positive integer and $a_1, a_2, \ldots, a_n$ are elements of an integral domain such that $a_1 a_2 \cdots a_n = 0$, show that at least one of the a's is zero.

5. If $\theta: R \to S$ is a homomorphism of the ring R into the ring S and $a \in R$, prove that $\theta(a^n) = \theta(a)^n$ for every positive integer n.

6. Prove that if a and b are elements of a ring, for every integer m (positive, negative, or zero),

$$m(a + b) = ma + mb.$$

7. Prove that if a and b are elements of a ring, for every integer m,

$$m(ab) = (ma)b = a(mb).$$

8. Prove that if a is an element of a ring, for all integers m and n,

$$ma + na = (m + n)a.$$

[Hint: Make a number of cases as follows: either m or n is zero; both m and n are positive; one of m, n is positive and the other negative; both m and n are negative.]

9. Use the fact that the set of positive integers is well-ordered to prove the following alternate form of the Induction Principle:

Suppose that there is associated with each positive integer n a statement S_n. Then S_n is true for every positive integer n provided the following hold:

(i) S_1 is true.

(ii) If k is a positive integer such that S_i is true for every positive integer $i < k$, then also S_k is true.

10. *Assume* the truth of the Induction Principle (3.10), and prove that the set of positive integers must be well-ordered. [Hint: Let S_n be the statement, "If a set of positive integers contains an element $\leq n$, then the set contains a least element."]

3.4 A CHARACTERIZATION OF THE RING OF INTEGERS

The purpose of this section is to prove the following theorem.

3.17 Theorem. *Let both D and D′ be ordered integral domains in which the set of positive elements is well-ordered. Then D and D′ are isomorphic.*

Since we are assuming that **Z** is an ordered integral domain in which the set of positive elements is well-ordered, this theorem will show that **Z** is the *only* ring with these properties (if we do not consider isomorphic rings as "different" rings).

As a first step in the proof, we shall prove a lemma. In the statement of this lemma, and henceforth whenever it is convenient to do so, we shall make use of the notation introduced in Definition 3.4 and let $\mathbf{Z}^+$ denote the set of all positive integers.

3.18 Lemma. *Let D be an ordered integral domain in which the set D^+ of positive elements is well-ordered. If e is the unity of D, then*

$$D^+ = \{me \mid m \in \mathbf{Z}^+\}$$

and

$$D = \{ne \mid n \in \mathbf{Z}\}.$$

Moreover, if $n_1, n_2 \in \mathbf{Z}$ such that $n_1 e = n_2 e$, then $n_1 = n_2$.

PROOF. We recall that $a \in D^+$ can also be expressed by writing $a > 0$. We have already observed that $e > 0$ and, in fact, that e is the least element of D^+. For each positive integer n, let S_n be the statement that $ne > 0$. Since, by definition, $1e = e$, we see that S_1 is true. If, now, k is a positive integer such that S_k is true, it follows from 3.5(i) that $(k+1)e = ke + e > 0$, and therefore that S_{k+1} is true. We have therefore proved by mathematical induction that $me > 0$ for every positive integer m. That is, $me \in D^+$ for every positive integer m. We now proceed to show that all elements of D^+ are of this form. Select any $d \in D^+$. If $d = e$ then d has the correct form. If $d \neq e$ then $d > e$ because e is the smallest element of D^+. Let U be the set of ***positive*** elements of the form $d - me$, m a positive integer. Then U is a nonempty set because $d - e$ is in U. By the well-ordering property, U has a least element u and $u = d - ne$ for some positive integer n. Since u is positive, we have $u = d - ne \geq e$ (again because e is the smallest positive element). This implies $d - (n+1)e \geq 0$. If equality does not hold, then the element $d - (n+1)e = u - e$ is smaller than u and also positive. But that would place $u - e$ in the set U and this would conflict with our choice of u as the smallest element of U. Hence the term equals 0 and $d = (n+1)e$, as required.

It is now easy to complete the proof of the first statement of the lemma. If $a \in D$, and $a \notin D^+$, then 3.4(iii) implies that $a = 0$ or $-a \in D^+$. If $a = 0$, then $a = 0 \cdot e$. If $-a \in D^+$, then by what we have just proved, $-a = m_2 e$ for some positive integer m_2. It follows that $a = (-m_2)e$, and so every element of D is of the form ne, where n is an integer (positive, negative, or zero).

Now suppose that $n_1, n_2 \in \mathbf{Z}$ such that $n_1 e = n_2 e$. If $n_1 \neq n_2$, we can assume that the notation is so chosen that $n_1 > n_2$. It follows that

$n_1 - n_2 > 0$ and, by the part of the lemma already proved, $(n_1 - n_2)e \in D^+$. Hence $(n_1 - n_2)e \neq 0$, or $n_1e \neq n_2e$. Thus the assumption that $n_1 \neq n_2$ leads to a contradiction, and we conclude that $n_1 = n_2$. This completes the proof of the lemma.

It is now easy to prove Theorem 3.17. If e and e' are the respective unities of D and of D', the lemma shows that

$$D = \{ne \mid n \in \mathbf{Z}\}$$

and

$$D' = \{ne' \mid n \in \mathbf{Z}\}.$$

Moreover, the last statement of the lemma asserts that the elements of D are *uniquely* expressible in the form ne, $n \in \mathbf{Z}$. Of course, the elements of D' are likewise uniquely expressible in the form ne', $n \in \mathbf{Z}$.

We now assert that the mapping $\theta: D \to D'$, defined by

$$\theta(ne) = ne', \qquad n \in \mathbf{Z},$$

is the desired isomorphism of D onto D'. By the uniqueness property just obtained, θ is a one-one mapping of D onto D'. Moreover, under this mapping we have

$$\begin{aligned}\theta(n_1e + n_2e) = \theta[(n_1 + n_2)e] &= (n_1 + n_2)e' \\ &= n_1e' + n_2e' = \theta(n_1e) + \theta(n_2e)\end{aligned}$$

and

$$\begin{aligned}\theta[(n_1e)(n_2e)] = \theta[(n_1n_2)e] &= (n_1n_2)e' \\ &= (n_1e')(n_2e') = [\theta(n_1e)][\theta(n_2e)].\end{aligned}$$

Hence, addition and multiplication are preserved and we indeed have an isomorphism. This completes the proof of the theorem.

EXERCISES

1. Let $D = 2^n$, $n \in \mathbf{Z}$ and let us define $2^n \oplus 2^m = 2^{n+m}$ and $2^n \odot 2^m = 2^{nm}$. Prove that with respect to the operations $\oplus$ and $\odot$ so defined, D is an ordered integral domain and that, in fact, D is isomorphic to $\mathbf{Z}$.
2. Let $R = \mathbf{Z} \oplus \mathbf{Z}$. Define an ordering on this ring by the rule $(a, b) < (c, d)$ if $a < c$ or if $a = c$ and $b < d$. (This is called the lexicographical or dictionary ordering on R since it resembles the rule for placing words in alphabetic order.) Describe the positive elements of R and show that R^+ does not have the property required of well-ordering. Give an explicit set of positive elements of R which does not have a least element.

3. Let R denote the ring of all real numbers with the usual ordering. Describe a nonempty set of positive elements of R which does not contain a least element. (One way to do this is to give a sequence $\{a_n\}$ of positive terms which has a zero limit.)

3.5 FIELDS

Let us make the following definition.

3.19 Definition. A commutative ring F with more than one element and having a unity is said to be a ***field*** if every nonzero element of F has a multiplicative inverse in F.

In view of Theorem 2.10, we know that every nonzero element of a field has a *unique* multiplicative inverse. As indicated in Section 2.4, we may denote the multiplicative inverse of a nonzero element r of a field F by r^{-1}. If 1 is the unity of F, r^{-1} is therefore the unique element of F such that

3.20 $$r \cdot r^{-1} = r^{-1} \cdot r = 1.$$

However, the commutative law of multiplication is required to hold in a field, and we shall henceforth use it without explicit mention. In particular, we may consider that r^{-1} is defined by the single equation $r \cdot r^{-1} = 1$.

We shall now prove the following result.

3.21 Theorem. *A field is necessarily an integral domain.*

PROOF. Suppose that r and s are elements of a field F such that $rs = 0$. If $r \neq 0$, r has a multiplicative inverse r^{-1} in F and it follows that

$$r^{-1}(rs) = (r^{-1}r)s = 1 \cdot s = s.$$

But also,

$$r^{-1}(rs) = r^{-1} \cdot 0 = 0.$$

Hence $s = 0$, and we have shown that $r = 0$ or $s = 0$. This proves that F has no nonzero divisors of zero and F therefore satisfies Definition 3.3 of an integral domain.

3.22 Corollary. *If R is a subring of a field F, and R contains the unity element of F, then R is an integral domain.*

PROOF. Since the operations of addition and multiplication of elements of R coincide with the same operations defined for F, it follows that R is a commutative ring with unity. Moreover, if r and s are elements of R with $rs = 0$, then, because r and s also belong to F, it follows that either $r = 0$ or $s = 0$. Hence R is an integral domain.

We shall see in the next section that there is a converse to this corollary. Every integral domain is a subring of a suitable field.

Although in the definition of a field we only required the existence of the multiplicative inverse of each nonzero element r, that is, the solvability of each equation of the form $rx = 1$, we can easily establish the following more general result.

3.23 Theorem. *If r and s are elements of a field F and $r \neq 0$, there exists a unique element y of F such that $ry = s$. Moreover, $y = r^{-1} \cdot s$.*

PROOF. It is clear that $r^{-1} \cdot s$ *is* a solution of this equation since

$$r(r^{-1} \cdot s) = (r \cdot r^{-1})s = 1 \cdot s = s.$$

To obtain the *uniqueness* of the solution, suppose that $ry_1 = s$ and $ry_2 = s$. Then $ry_1 = ry_2$ and, since $r \neq 0$, the cancellation law of multiplication shows that $y_1 = y_2$.

Let us now give a few examples of fields. It is to be understood that the usual definitions of addition and multiplication are implied.

Example 1: The set of all rational numbers; that is, all numbers of the form a/b, where $a, b \in \mathbf{Z}$ with $b \neq 0$.

Example 2: The set of all real numbers.

Example 3: The set of all real numbers of the form $x + y\sqrt{2}$, where x and y are rational numbers. What is the multiplicative inverse of each nonzero element?

Example 4: The set of all real numbers of the form $u + v\sqrt{3}$, where u and v are elements of the field of the preceding example.

To verify that the sets in Examples 3 and 4 are indeed fields, it is necessary to check that each set is closed under addition and multiplication. Since the existence of the unity element is clear, it then follows that each set is a subring of the field of real numbers. By Corollary 3.22, each set is an integral domain. All that remains is to show that each nonzero element of the set has a multiplicative inverse in the set. Of course each nonzero element has an inverse in the field of real numbers; it must be shown that the inverse actually lies in the set in question. This amounts to the process of "rationalizing the denominator" which is discussed in elementary arithmetic. This is left as an exercise for the reader.

Subfields and Extensions

It is often useful to consider a given field in relation to a larger field containing it or in relation to fields contained inside of it. The following definitions are convenient.

3.24 Definition. A subring F' of a field F which is itself a field is called a ***subfield*** of F. If F' is a subfield of F, F is called an ***extension*** of F'.

The fields in Examples 2, 3, 4 are all extensions of **Q** and all are subfields of the field of real numbers.

3.6 FIELD OF QUOTIENTS

We now change our point of view as follows. Instead of studying properties of a given field, let us see how we can start with a given integral domain D and *construct* a field which contains D. This is our first example of an important algebraic problem which may be stated in a general way as follows. Given an algebraic system U which does not have some specific property, construct a larger system V which contains U and which does have the property in question. Naturally, this will not always be possible, but it will be possible in many interesting cases. At present, we begin with an integral domain D in which some nonzero elements do not have a multiplicative inverse, and construct a field F which contains D and, of course, every nonzero element of F does have a multiplicative inverse. To be technically correct, it should be mentioned that the construction will produce a field F which contains a subring isomorphic with D. If we treat isomorphic rings as equal, then it will be correct to say that F contains D.

In order to motivate the ideas used in the construction to follow, we first provide some familiar notation which will be suggestive later on.

Let F be a field with unity 1, and t a nonzero element of F. We have introduced the symbol t^{-1} to designate the multiplicative inverse of t, and have found that if $s \in F$, the unique element x of F such that $tx = s$ is given by $x = t^{-1}s$. In accordance with familiar usage, we shall also designate this element $t^{-1}s$ by $\frac{s}{t}$ or by s/t. In particular, we have $t^{-1} = 1/t$.

Suppose, now, that v is also a nonzero element of F. Since $(tv)(v^{-1}t^{-1}) = 1$, the multiplicative inverse of tv is $v^{-1}t^{-1}$, that is,

3.25
$$(tv)^{-1} = v^{-1}t^{-1}.$$

It is now easy to see that

3.26
$$\frac{sv}{tv} = \frac{s}{t}$$

from the following calculation:

$$\frac{sv}{tv} = (tv)^{-1}(sv) = v^{-1}t^{-1}sv = t^{-1}s = \frac{s}{t}.$$

As a generalization of 3.26, let s and u be arbitrary elements of F, and t and v arbitrary nonzero elements of F. Then we assert that

3.27
$$\frac{s}{t} = \frac{u}{v} \quad \text{if and only if} \quad sv = tu.$$

Suppose, first, that $s/t = u/v$, that is, that $t^{-1}s = v^{-1}u$. Multiplication by tv yields $sv = tu$. Conversely, if $sv = tu$, multiplication by $t^{-1}v^{-1}$ shows that $t^{-1}s = v^{-1}u$ or, otherwise expressed, that $s/t = u/v$.

The following are also easy to establish and will be left as exercises:

3.28

(i)
$$\frac{s}{t} + \frac{u}{v} = \frac{sv + tu}{tv},$$

(ii)
$$\frac{s}{t} \cdot \frac{u}{v} = \frac{su}{tv}.$$

Now we turn to the construction of a field containing the given integral domain D. Let S denote the set of all ordered pairs (a, b) with $a, b \in D$ and $b \neq 0$; that is,

$$S = \{(a, b) \mid a, b \in D, b \neq 0\}.$$

Our next step will be suggested by thinking of (a, b) as the familiar a/b, but we use an unfamiliar notation in order to clarify the logical procedure and to avoid using any property until we have actually proved it. Keep in mind that the symbol a/b has no meaning until there is some field containing a and b. Another point to be considered is that $a/b = au/bu$ for any nonzero u in a field, so that somehow the pairs (a, b) and (au, bu) should represent the same element, or, at least, these two pairs should be related. Accordingly we define a relation on the set S. This will be an equivalence relation as defined in Chapter 1. For elements (a, b) and (c, d) of S, we define $(a, b) \sim (c, d)$ to mean that $ad = bc$. In order to show that "$\sim$" is an equivalence relation, the three conditions of Definition 1.8 must be verified. The reflexive and symmetric properties are easily checked, so we prove the transitive property. Suppose that $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$; it is necessary to show that $(a, b) \sim (e, f)$. Since $(a, b) \sim (c, d)$, we have $ad = bc$; similarly, we have $cf = de$. Multiplication of these equations by f and b, respectively, yields $adf = bcf$ and $bcf = bde$. Thus $adf = bde$ and, since $d \neq 0$, it follows that $af = be$; this means $(a, b) \sim (e, f)$.

Now that we have an equivalence relation defined on S, we may consider the equivalence classes. Our previous usage would dictate that the equivalence class containing (a, b) would be denoted as $[(a, b)]$. However, we shall use the simpler notation $[a, b]$ to denote this class. In the sequel it is important to keep in mind that $[a, b] = [u, v]$ if and only if $(a, b) \sim (u, v)$, that is, if and only if, $av = bu$. The equivalence class $[a, b]$ may be expressed as

$$[a, b] = \{(x, y) \mid (x, y) \in S, xb = ay\}.$$

Now let F denote the set of all equivalence classes:

$$F = \{[a, b] \mid (a, b) \in S\}.$$

We shall define addition and multiplication on the set F and then verify that we have obtained a field. By keeping in mind that $[a, b]$ should eventually represent the fraction a/b, Equations 3.28 suggest the definitions we need.

We define addition and multiplication of equivalence classes as follows:

3.29(i) $$[a, b] + [c, d] = [ad + bc, bd]$$

and

3.29(ii) $$[a, b] \cdot [c, d] = [ac, bd].$$

First, we observe that since (a, b) and (c, d) are elements of S, we have $b \neq 0$ and $d \neq 0$. Hence $bd \neq 0$ and so the elements on the right side of Equations 3.29(i) and (ii) are indeed equivalence classes. However, there is an important point which must yet be considered before we may logically conclude that we have defined two operations on F. Consider how addition is carried out using the formula in Equation 3.29(i). Given the equivalence classes $X = [a, b]$ and $Y = [c, d]$, their sum, $X + Y$, is determined as the equivalence class containing a certain element of S which is computed in terms of a, b, c, d. But the equivalence class $[a, b]$ may contain many elements, and as such it may be represented by many different pairs. Suppose that $[a, b] = [a_1, b_1]$ and $[c, d] = [c_1, d_1]$. Then our formula for addition implies that the sum of the two classes should yield

$$X + Y = [ad + bc, bd]$$

and also

$$X + Y = [a_1 d_1 + b_1 c_1, b_1 c_1].$$

Clearly the left sides of these two equations are equal; if our rule for addition is a good one, then the right sides of these two equations must also be equal. This problem will occur again; the term *well-defined* is used in this context. It is necessary to show that addition given by Equation 3.29 is well-defined. Of course, it must also be shown that multiplication is well-defined. Suppose, then, that

3.30 $$[a,b]=[a_1,b_1] \quad \text{and} \quad [c,d]=[c_1,d_1].$$

In order to show that addition of equivalence classes is well-defined by 3.29 we must show that necessarily

$$[a,b]+[c,d]=[a_1,b_1]+[c_1,d_1],$$

that is, that

3.31 $$[ad+bc,bd]=[a_1d_1+b_1c_1,b_1d_1].$$

From 3.30, we have that $ab_1=ba_1$ and that $cd_1=dc_1$. If we multiply the first of these equations by dd_1, the second by bb_1, and add the corresponding members, it follows that

$$(ad+bc)b_1d_1=bd(a_1d_1+b_1c_1).$$

However, this implies 3.31, and therefore addition of equivalence classes is well-defined by 3.29(i). The proof that multiplication is well-defined by 3.29(ii) will be left as an exercise.

We may now state the following theorem.

3.32 Theorem. *Let F denote the set of all equivalence classes of S relative to the equivalence relation "~." Then with respect to the operations of addition and multiplication on F defined by 3.29* (i) *and* (ii), $a \in D$, *is a subring D′ of F; and the mapping* $\theta: D \to D'$, *defined by*

F is a field. Moreover, the set of all elements of F of the form $[a,1]$,

$$\theta(a)=[a,1], \qquad a \in D,$$

is an isomorphism of D onto D′.

PROOF. The commutative laws of addition and multiplication, as well as the associative law of multiplication, are almost abvious, and we omit the details. The associative law of addition may be verified by the following straightforward calculation. Let $[a,b]$, $[c,d]$, and $[e,f]$ be elements of F. Then

$$\begin{aligned}([a,b]+[c,d])+[e,f]&=[ad+bc,bd]+[e,f]\\&=[adf+bcf+bde,bdf]\end{aligned}$$

and

$$\begin{aligned}[a,b] + ([c,d] + [e,f]) &= [a,b] + [cf + de, df] \\ &= [adf + bcf + bde, bdf],\end{aligned}$$

and we therefore have

$$([a,b] + [c,d]) + [e,f] = [a,b] + ([c,d] + [e,f]).$$

Since $[0,1] + [a,b] = [a,b]$ and $[1,1]\cdot[a,b] = [a,b]$, it follows that $[0,1]$ is the zero and $[1,1]$ the unity of F. However, if d is a nonzero integer, we have $[d,d] = [1,1]$ and, similarly, $[0,1] = [0,d]$. Hence the unity is $[d,d]$ and the zero is $[0,d]$ for *any* nonzero integer d. We may also observe that $[a,b] = [0,1]$ if and only if $a = 0$, and to say, therefore, that $[a,b]$ is a nonzero element of F is to say that $a \neq 0$.

Since $[a,b] + [-a,b] = [0,b^2]$ and $[0,b^2]$ is the zero of F, it follows that the additive inverse of $[a,b]$ is $[-a,b]$, that is, we have $-[a,b] = [-a,b]$, and each element of F has an additive inverse.

One of the distributive laws is a consequence of the following calculations in which, at one point, we make use of the fact that $[b,b]$ is the unity of F:

$$\begin{aligned}[a,b]([c,d] + [e,f]) &= [a,b]\cdot[cf + de, df] \\ &= [acf + ade, bdf], \\ [a,b]\cdot[c,d] + [a,b]\cdot[e,f] &= [ac, bd] + [ae, bf] \\ &= [acbf + bdae, b^2df] \\ &= [acf + ade, bdf]\cdot[b,b] \\ &= [acf + ade, bdf].\end{aligned}$$

The other distributive law is an immediate consequence of this one since multiplication is commutative.

Up to this point we have proved that F is a commutative ring with unity. To prove that F is a field, there remains only to show that every nonzero element of F has a multiplicative inverse in F. If $[a,b]$ is a nonzero element of F, then $a \neq 0$ as well as $b \neq 0$, and it is clear that $[b,a] \in F$. Moreover,

$$[a,b]\cdot[b,a] = [ab, ab] = [1,1],$$

and the multiplicative inverse of $[a,b]$ is $[b,a]$. That is, if $[a,b]$ is a nonzero element of F, then $[a,b]^{-1} = [b,a]$. This completes the proof that F is a field.

Now let D' be the set of elements of F of the form $[a,1]$, $a \in D$, and consider the mapping $\theta: D \to D'$ defined by $\theta a = [a,1]$, $a \in D$. This is clearly an onto mapping and it is also a one-one mapping since $[a,1] = [b,1]$ implies that $a = b$. Moreover, for $a, b \in D$, we have

$$\theta(a + b) = [a + b, 1] = [a,1] + [b,1] = \theta(a) + \theta(b)$$

and

$$\theta(ab) = [ab, 1] = [a, 1]\cdot[b, 1] = \theta(a)\theta(b).$$

Thus θ is an isomorphism of D onto D', and the theorem is established.

Since the subring D of F is isomorphic to D, we shall henceforth find it convenient to identify D' with D and, as a matter of notation, write simply a to designate the element $[a, 1]$ of F. We may then consider that the field F actually contains the ring D.

As a further simplification of notation, let us observe that

$$[a, b] = [a, 1]\cdot[1, b] = [a, 1]\cdot[b, 1]^{-1},$$

and hence we are justified in writing $a\cdot b^{-1}$ or a/b for the element $[a, b]$ of F. This justifies the "familiar" notation introduced earlier in this section. The field F constructed using the integral domain D is called the *field of quotients* of D.

In the special case in which D is the ring of integers, usually denoted by **Z**, the field of quotients is the *field of rational numbers* and is denoted by **Q**.

In the notation which we have finally introduced, the field **Q** consists of all numbers of the form a/b, where a and b are integers with $b \neq 0$, addition and multiplication being defined in the usual way (3.29(i) and (ii)).

Let us emphasize the meaning of the notation we have introduced by considering, for example, the rational number $1/2$. We are writing $1/2$ for the equivalence class $[1, 2]$ used above. Now $[1, 2] = [c, d]$ if and only if $d = 2c$, so we see that $1/2$ represents the equivalence class consisting of all ordered pairs of the form $(c, 2c)$, where c is a nonzero integer. Moreover, for example, $1/2 = 3/6$ simply because by our definition of equivalence, $(1, 2) \sim (3, 6)$ and therefore $[1, 2] = [3, 6]$.

Since $(-a)/b = a/(-b)$, we see that every rational number can be written in the form c/d, where $d > 0$. Moreover, if the integers c and d have a common nonzero factor k, so that $c = c_1k$ and $d = d_1k$, then $c/d = c_1/d_1$. It follows that every nonzero rational number r can be written in the form a/b, where a and b have highest common factor equal to 1 and $b > 0$. If r is expressed in this form, it is sometimes said that r is expressed *in lowest terms.*

EXERCISES

1. Suppose the integral domain D is already a field. Show that the construction of the field of quotients of D given in Theorem 3.32 produces a field isomorphic to D.
2. Examine the construction of the field of quotients of the integral domain D and then explain why it is necessary that D be an integral domain, and not just a commutative ring with unity.

3. Let n be a nonzero positive integer and let D be the subset of all rational numbers of the form a/b where $a, b \in \mathbf{Z}$, and $b = n^k$, for some nonnegative integer k. Show that D is an integral domain and that its field of quotients is isomorphic to the field of rational numbers.

3.7 A FEW PROPERTIES OF THE FIELD OF RATIONAL NUMBERS

We have defined in Section 3.2 what we mean by an ordered integral domain. Since a field is necessarily an integral domain, by an *ordered field* we shall naturally mean a field which is an ordered integral domain. We shall now prove the following result.

3.33 Theorem. *Let $\mathbf{Q}^+$ denote the set of all rational numbers a/b, where a and b are integers such that $ab > 0$. Then $\mathbf{Q}^+$ has the properties 3.4 which define an ordered integral domain, and therefore the field $\mathbf{Q}$ is an ordered field whose positive elements are the elements of $\mathbf{Q}^+$.*

We may point out that when we write $ab > 0$, we mean that ab is a positive *integer* and we are only making use of the fact that $\mathbf{Z}$ is an ordered integral domain.

PROOF. First, we need to show that the definition of an element of $\mathbf{Q}^+$ does not depend upon the particular representation of a rational number. That is, we need to show that if $a/b = c/d$ and $ab > 0$, then also $cd > 0$. This follows from the observation that $a/b = c/d$ means that $ad = bc$ and $ab > 0$ implies that either a and b are both positive or they are both negative. The same must therefore be true of c and d; hence also $cd > 0$.

Now let us show (3.4(i)) that the set $\mathbf{Q}^+$ is closed under addition. Let a/b and c/d be elements of $\mathbf{Q}^+$, and therefore $ab > 0$ and $cd > 0$. Then

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd},$$

and we wish to show that

$$(ad + bc)bd = abd^2 + cdb^2 > 0.$$

However, this inequality follows easily from the following known inequalities: $ab > 0$, $cd > 0$, $b^2 > 0$, and $d^2 > 0$.

It is trivial that $\mathbf{Q}^+$ is closed under multiplication (3.4(ii)). Moreover, if a/b is a nonzero rational number, then either $ab > 0$ or $ab < 0$. It follows that for every rational number a/b, exactly one of the following holds (3.4(iii)):

$$\frac{a}{b}=0, \quad \frac{a}{b}>0, \quad -\frac{a}{b}>0.$$

Hence $\mathbf{Q}^+$ has the three required properties, and the field $\mathbf{Q}$ of rational numbers is ordered.

It will be observed that what we have done is to make use of the known ordering of the integers to establish an ordering of the rational numbers. Inasmuch as we have identified the integer a with the rational number $a/1$, it is clear that a is a positive integer if and only if a is a positive rational number. In other words, our ordering of the rational numbers is an *extension* of the previous ordering of the integers.

In view of Theorem 3.33, we can introduce inequalities involving rational numbers in the usual way. That is, if $r, s \in \mathbf{Q}$, we write $r > s$ (or $s < r$) to mean that $r - s \in \mathbf{Q}^+$, and so on. We now have available all the usual properties (3.6) of inequalities for rational numbers. In the future we shall make use of these properties without specific reference.

The following is a significant property of the rational numbers.

3.34 Theorem. *Between any two distinct rational numbers there is another rational number.*

PROOF. Suppose that $r, s \in \mathbf{Q}$ with $r < s$. The theorem will be established by showing that

$$r < \frac{r+s}{2} < s,$$

and hence that $(r+s)/2$ is a rational number between r and s. Since $r < s$, we have $r + r < r + s$, or $2r < r + s$. Now multiplying this last inequality by the positive rational number $1/2$, we obtain $r < (r+s)/2$. In a similar manner, it can be shown that $(r+s)/2 < s$, and we omit the details.

The property of the rational numbers stated in the preceding theorem is often expressed by saying that the rational numbers are *dense*. We shall now prove in the following theorem another simple, but important, property of the rational numbers.

3.35 Theorem (ARCHIMEDEAN PROPERTY). *If r and s are any positive rational numbers, there exists a positive integer n such that $nr > s$.*

PROOF. Let $r = a/b, s = c/d$, where a, b, c, and d are positive integers. If n is a positive integer, then $n(a/b) > c/d$ if and only if $n(ad) > bc$. We now assert that this last inequality is necessarily satisfied if we choose $n = 2bc$. For $ad \geq 1$, and therefore $2ad > 1$. Multiplying this inequality by the positive integer bc shows that $2adbc > bc$. Hence, $n = 2bc$ certainly satisfies our requirement. Of course, we do not mean to imply that this is necessarily the smallest possible choice of n.

3.8 THE FIELD OF REAL NUMBERS

The rational numbers are sufficient for use in all simple applications of mathematics to physical problems. For example, measurements are usually given to a certain number of decimal places, and any finite decimal is a rational number. However, from a theoretical point of view, the system of rational numbers is entirely inadequate. The Pythagoreans made this discovery about 500 B.C. and were shocked by it. Consider, for example, an isosceles right triangle whose legs are 1 unit in length. Then, by the Pythagorean theorem, the hypotenuse has length $\sqrt{2}$; and from this geometrical consideration it appears that there must exist a "number" $\sqrt{2}$, although we will see in Section 4.9 that it cannot be a rational number.

The inherent difficulty in extending the field of rational numbers to the field of real numbers is perhaps indicated by the fact that a satisfactory theory of the real numbers was not obtained until the latter half of the nineteenth century. Although other men also made contributions to the theory, it is usually attributed to the German mathematicians Dedekind (1831–1916) and Cantor (1845–1918). We shall not present here the work of either of these men but shall presently state without proof the fundamental theorem which each of them essentially proved and by quite different methods. In order to do this, we must first make a few preliminary definitions.

So far, the only ordered field which we have studied is the field Q of rational numbers. However, for the moment, suppose that F is an arbitrary ordered field and let us make the following definition.

3.36 Definition. Let S be a set of elements of an ordered field F. If there exists an element b of F such that $x \leq b$ for every element x of S, then b is called an ***upper bound*** of the set S in F.

As an example, the set $S_1 = \{\frac{1}{2}, 1, 2\}$ of elements of **Q** has an upper bound 2. Also $\frac{5}{2}$ is an upper bound of this set, as is 117, and so on. Thus, if a set has an upper bound, it has many upper bounds. Clearly, the set $\mathbf{Z}^+$ of all positive integers does not have an upper bound in **Q**. As another example, consider the set

$$S_2 = \{a \mid a \in \mathbf{Q}, a > 0, a^2 < 2\}.$$

Then S_2 has upper bounds in **Q**, one of them being 3.

3.37 Definition. Let S be a set of elements of an ordered field F. If there exists an upper bound c of S in F such that no smaller element of F is an upper bound of S, then c is called the ***least upper bound*** (l.u.b.) of S in F.

It follows from this definition that if a set S has a l.u.b., it is unique. Moreover, if c is the l.u.b. of the set S in F and $d \in F$ such that $d < c$, then there must exist an element s of S such that $s > d$ since, otherwise, d would be an upper bound of S less than the least upper bound.

For the set S_1 exhibited above, the element 2 of S_1 is clearly the l.u.b. of S_1 in **Q**. However, for the set S_2 the situation is not quite so obvious. Although we shall not give the details, it is true and should perhaps not be surprising that there exists no rational number which is the l.u.b. of the set S_2. That is, if $c \in \mathbf{Q}$ is an upper bound of S_2, there exists $d \in \mathbf{Q}$ such that $d < c$ is also an upper bound of S_2. Therefore S_2 has no l.u.b. in **Q**. Thus we have an example of a set of elements of **Q** which has upper bounds in **Q** but no l.u.b. in **Q**. In the field of real numbers, whose existence is asserted in the next theorem, this situation cannot arise. In fact, it is the existence of an ordered field with this property which may be considered to be the principal contribution of Dedekind and Cantor to the subject. Let us state this result as the following theorem.

3.38 Theorem. *There exists a field* **R**, *called the field of real numbers, with the following properties:*

(*i*) **R** *is an extension of the field* **Q** *of rational numbers. Moreover,* **R** *is an ordered field and* $\mathbf{Q}^+ \subset \mathbf{R}^+$.

(*ii*) *If S is a nonempty set of elements of* **R** *which has an upper bound in* **R**, *it has a l.u.b. in* **R**.

The elements of **R** are called *real numbers*. An element of **R** which is not an element of **Q** is called an *irrational* number. Considered as elements of **R**, the set S_2 defined above has a l.u.b. (by 3.38(ii)) and this l.u.b. we may *define* to be the number $\sqrt{2}$. Clearly, $\sqrt{2}$ is an irrational number and it is the fact that this number is not an element of **Q** which prevents S_2 from having a l.u.b. *in the field* **Q**.

The fact that $\mathbf{Q}^+ \subset \mathbf{R}^+$ is sometimes expressed by saying that the ordering of **R** is an extension of the ordering of **Q**. That is, a rational number is a positive rational number if and only if, considered as a real number, it is a positive real number. A similar situation arose when we passed from the integers to the rational numbers.

3.9 SOME PROPERTIES OF THE FIELD OF REAL NUMBERS

In this section we shall prove two fundamental properties of real numbers and state one additional property without proof. Of course, our proofs will be based on the assumed properties (i) and (ii) of Theorem 3.38.

Throughout the rest of this book we shall continue to denote the field of real numbers by **R** and the set of positive real numbers by $\mathbf{R}^+$.

3.39 Theorem (Archimedean Property). *If* $a, b \in \mathbf{R}^+$, *there exists a positive integer* n *such that* $na > b$.

PROOF. Let us assume that $ka \leq b$ for every positive integer k, and seek a contradiction. Another way of stating this assumption is to assert that b is an upper bound of the set $S = \{ka \mid k \in \mathbf{Z}^+\}$. Since this set has an upper bound, by 3.38(ii) it has a l.u.b., say c. Now $c - a < c$ and therefore $c - a$ is not an upper bound of the set S. This implies that there exists an element la of S, $l \in \mathbf{Z}^+$, such that $la > c - a$. It follows that $(l + 1)a > c$ and since $(l + 1)a \in S$, we have a contradiction of the fact that c is the l.u.b. of the set S. The proof is therefore complete.

It was shown in Section 3.7 that between any two distinct rational numbers there is another rational number. A generalization of the result is given in the following theorem.

3.40 Theorem. *If* $a, b \in \mathbf{R}$ *with* $a < b$, *there exists a rational number* m/n $(m, n \in \mathbf{Z})$ *such that*

$$a < \frac{m}{n} < b.$$

PROOF. For simplicity, we shall assume that $a > 0$ and leave the rest of the proof as an exercise.

Since $b - a > 0$, by the preceding theorem there exists $n \in \mathbf{Z}^+$ such that $n(b - a) > 1$. Let n be some such fixed integer. Again applying the preceding theorem to the real numbers 1 and na, there exists $m \in \mathbf{Z}^+$ such that $m > na$, and let m be the *least* positive integer with this property. Now $m > na$ implies that $a < m/n$ and we proceed to complete the proof by showing that also $m/n < b$ or, equivalently, that $m < nb$. Suppose that $m \geq nb$. Since $n(b - a) > 1$, we have $m \geq nb > na + 1$. Thus $m > 1$ and $(m - 1) \in \mathbf{Z}^+$ such that $(m - 1) > na$. Since $m - 1 < m$, this violates our choice of m as the least positive integer which is greater than na. Our assumption that $m \geq nb$ has led to a contradiction, and we conclude that $m < nb$. This completes the proof.

In particular, this theorem tells us that between any two irrational numbers there is a rational number. It is also true that between any two rational numbers there is an irrational number. (See Exercise 2 below.) Thus the rational and irrational numbers are very closely intertwined.

Although it is true that all the properties of the real numbers can be established using only the properties (i) and (ii) of Theorem 3.38 we shall give no further proofs in this book. However, let us conclude this brief discussion of the real numbers by stating without proof the following familiar and important result.

3.41 Theorem. *For each positive real number a and each positive integer n, there exists exactly one positive real number x such that $x^n = a$.*

The real number x whose existence is asserted by this theorem may be called the *principal* nth root of a and designated by the familiar notation $a^{1/n}$ or by $\sqrt[n]{a}$.

EXERCISES

1. Define lower bound and greatest lower bound of a set of elements of an ordered field. Prove that if a nonempty set of elements of $\mathbf{R}$ has a lower bound, it has a greatest lower bound in $\mathbf{R}$. [Hint: Consider the additive inverses of elements of the set.]
2. Prove that if $a, b \in \mathbf{R}$ with $a < b$, then

$$a < a + \frac{b - a}{\sqrt{2}} < b.$$

Hence prove that between any two distinct rational numbers there is an irrational number.

3. Complete the proof of Theorem 3.40 by showing that the stated result holds also for the case in which $a \leq 0$.
4. Let S_1 and S_2 be nonempty sets of real numbers having, respectively, b_1 and b_2 as least upper bounds. If $S_3 = \{s_1 + s_2 \mid s_1 \in S_1, s_2 \in S_2\}$, prove that $b_1 + b_2$ is the least upper bound of the set S_3.

3.10 THE FIELD OF COMPLEX NUMBERS

In order to construct the field of complex numbers, we begin by considering ordered pairs (a, b) of *real* numbers. Our definitions of addition and multiplication will be motivated by the formal properties of expressions of the form $a + bi$, where $i^2 = -1$. However, we are not justified in assuming that there *is* a "number" whose square is -1 until we have constructed a field which has an element with this property. Accordingly, as in the case of the construction of the rational numbers, we begin with an unfamiliar notation in order to avoid using any property until we have established it. We may remind the reader that the equal sign is being used in the sense of identity, that is, $(a, b) = (c, d)$ means that $a = c$ and $b = d$.

We proceed to prove the following theorem, which establishes the existence of the field we shall presently call the field of complex numbers.

3.42 Theorem. *Let* **C** *be the set of all ordered pairs* (a, b) *of elements of the field* **R** *of real numbers, and let us define operations of addition and multiplication on* **C** *as follows:*

3.43 $$(a, b) + (c, d) = (a + c, b + d),$$

3.44 $$(a, b)(c, d) = (ac - bd, ad + bc).$$

Then **C** *is a field with respect to these definitions of addition and multiplication. Moreover, the set of all elements of* **C** *of the form* $(a, 0)$, $a \in$ **R**, *is a subfield of* **C** *which is isomorphic to the field* **R**.

PROOF. The required properties of addition are almost obvious. From 3.43, it follows that addition is commutative and associative, that $(0, 0)$ is the zero of **C**, and that the additive inverse of (a, b) is $(-a, -b)$.

The associative law of multiplication is a consequence of the following straightforward calculations:

$$\begin{aligned}((a, b)(c, d))(e, f) &= (ac - bd, ad + bc)(e, f)\\ &= (ace - bde - adf - bcf, acf - bdf + ade + bce),\\ (a, b)((c, d)(e, f)) &= (a, b)(ce - df, cf + de)\\ &= (ace - adf - bcf - bde, acf + ade + bce - bdf),\end{aligned}$$

and these turn out to be equal elements of **C**.

Next, let us verify one of the distributive laws as follows:

$$\begin{aligned}(a, b)((c, d) + (e, f)) &= (a, b)(c + e, d + f)\\ &= (ac + ae - bd - bf, ad + af + bc + be),\\ (a, b)(c, d) + (a, b)(e, f) &= (ac - bd, ad + bc) + (ae - bf, af + be)\\ &= (ac - bd + ae - bf, ad + bc + af + be),\end{aligned}$$

and again we have equal elements of **C**. The other distributive law follows from this one as soon as we show that multiplication is commutative, and the commutativity of multiplication follows easily from 3.44. For, by interchanging (a,b) and (c,d) in 3.44, we see that

$$(c,d)(a,b) = (ca - db, cb + da),$$

and the right side of this equation is equal to the right side of 3.44. Hence,

$$(a,b)(c,d) = (c,d)(a,b).$$

We have now proved that **C** is a commutative ring, and it is easily verified that it has the unity $(1,0)$. To show that **C** is a field, we need only show that each nonzero element (a,b) of **C** has a multiplicative inverse in **C**. Since the zero is $(0,0)$, to say that (a,b) is not the zero of **C** is to say that a and b are not both equal to zero. Since a is an element of the ordered field **R**, we know that if $a \neq 0$, then $a^2 > 0$. Similarly, if $b \neq 0$, we have $b^2 > 0$. It follows that if (a,b) is not the zero of **C**, then necessarily $a^2 + b^2 > 0$ and, in particular, $a^2 + b^2 \neq 0$. Hence,

$$\left(\frac{a}{a^2+b^2}, \frac{-b}{a^2+b^2}\right)$$

is an element of **C** and it may be verified by direct calculation (using 3.44) that

$$(a,b)\left(\frac{a}{a^2+b^2}, \frac{-b}{a^2+b^2}\right) = (1,0).$$

We have therefore shown that every nonzero element of **C** has a multiplicative inverse in **C**, and hence we have proved that **C** is a field.

To complete the proof of the theorem, let **R**′ be the set of all elements of **C** of the form $(a,0)$, $a \in \mathbf{R}$. Then the mapping $\theta: \mathbf{R}' \to \mathbf{R}$ defined by $\theta(a,0) = a$, $a \in \mathbf{R}$, is a one-one mapping of **R**′ onto **R**. Moreover,

$$\theta[(a,0) + (b,0)] = \theta(a+b,0) = a + b = \theta(a,0) + \theta(b,0)$$

and

$$\theta[(a,0)(b,0)] = (ab,0) = ab = [\theta(a,0)][\theta(b,0)].$$

Hence, the operations of addition and multiplication are preserved under this mapping, and the mapping therefore defines an isomorphism of **R**′ onto **R**. This completes the proof of the theorem.

An element of the field which we have constructed is called a *complex number*, and **C** is called the *field of complex numbers*.

We shall henceforth adopt a more familiar notation by identifying **R**′ with **R**, that is, we shall write a in place of $(a,0)$ and consider that the field **C** of complex numbers actually contains the field **R** of real numbers. Also, for simplicity of notation, as well as for historical reasons, we shall use the symbol i to designate the particular element $(0, 1)$ of **C**. Since $(0, 1)^2 = (-1, 0)$, in our new notation we have $i^2 = -1$. Now it is easily verified that

$$(a, 0) + (b, 0)(0, 1) = (a, b)$$

and, using the notation we have introduced, it follows that $a + bi = (a, b)$. Accordingly, in the future we shall write $a + bi$ in place of (a, b). In this notation, the product 3.44 of two elements of **C** may be expressed in the following form:

$$(a + bi)(c + di) = ac - bd + (ad + bc)i. \tag{3.45}$$

Of course, the right side of 3.45 may be obtained from the left by multiplying out with the aid of the usual distributive, associative, and commutative laws, and replacing i^2 by -1.

We have now extended the field of real numbers to the field of complex numbers. It should be pointed out, however, that one familiar property of the field of rational numbers and of the field of real numbers does not carry over to the field of complex numbers.

3.46 Theorem. *The field* **C** *of complex numbers is not an ordered field.*

PROOF. By this statement we mean that there does not exist any set $\mathbf{C}^+$ of elements of **C** having the properties (3.4) required for **C** to be an ordered field. This fact is a consequence of the following observations. If **C** were ordered, 3.6(v) would show that the square of every nonzero element would be positive; in particular, both i^2 and 1 would be positive. Then -1 would be negative, and we have a contradiction since $i^2 = -1$.

The fact that **C** is not ordered means that inequalities cannot be used between complex numbers. In other words, it is meaningless to speak of one complex number as being greater or less than another.

Throughout the rest of this book we shall continue to denote the field of complex numbers by **C**.

The Conjugate of a Complex Number

Let us make the following definition.

3.47 Definition. If $u = a + bi \in \mathbf{C}$, we define the ***conjugate*** of u to be the element u^* of **C** given by $u^* = a - bi$.[†]

[†] Historically, the usual notation for the conjugate of a complex number z is $\bar{z}$ instead of z^*. The present notation has been adopted only because it makes it easier to print such expressions as the conjugate of the sum of two or more complex numbers.

As examples, we have $(1 + 7i)^* = 1 - 7i$, $(2 - 2i)^* = 2 + 2i$, $4^* = 4$, and so on.

Now the mapping $\alpha: \mathbf{C} \to \mathbf{C}$ defined by $\alpha(u) = u^*$, $u \in \mathbf{C}$, is a one-one mapping of **C** onto **C**. We proceed to show that the operations of addition and multiplication are preserved under this mapping. Let $u = a + bi$ and $v = c + di$ be elements of **C**. Then

$$\begin{aligned}\alpha(u + v) &= (u + v)^* = [(a + c) + (b + d)i]^* = a + c - (b + d)i \\ &= (a - bi) + (c - di) = u^* + v^* = \alpha(u) + \alpha(v)\end{aligned}$$

and

$$\begin{aligned}\alpha(uv) &= (uv)^* = [ac - bd + (ad + bc)i]^* = ac - bd - (ad + bc)i \\ &= (a - bi)(c - di) = u^*v^* = \alpha(u)\alpha(v).\end{aligned}$$

Since the operations of addition and multiplication are preserved under the mapping α, it follows that this is an isomorphism of the field **C** onto itself. Such an isomorphism is frequently called an *automorphism*.

In working with complex numbers, the concept of conjugate plays an important role. A number of simple, but significant, properties are presented in Exercise 2 which follows.

EXERCISES

1. Find the multiplicative inverse of the nonzero element (a, b) of **C** by assuming that r and s are real numbers such that $(a, b)(r, s) = (1, 0)$, and solving for r and s.
2. Prove each of the following:
   (i) If $u \in \mathbf{C}$, then $uu^* \in \mathbf{R}$ and $u + u^* \in \mathbf{R}$; moreover, if $u \neq 0$, then $uu^* > 0$.
   (ii) If $u \in \mathbf{C}$, then $(u^*)^* = u$.
   (iii) If $u \in \mathbf{C}$ and $u \neq 0$, then $(u^{-1})^* = (u^*)^{-1}$.
   (iv) If $u \in \mathbf{C}$, then $u = u^*$ if and only if $u \in \mathbf{R}$.
   (v) If $u \in \mathbf{C}$ and n is a positive integer, then $(u^n)^* = (u^*)^n$.
3. If $\alpha: \mathbf{C} \to \mathbf{C}$ is the isomorphism defined above, part (iv) of the preceding exercise shows that $\alpha(u) = u$ if and only if $u \in \mathbf{R}$. Prove that if $\phi: \mathbf{C} \to \mathbf{C}$ is an isomorphism of **C** onto **C** with the property that $\phi(a) = a$ for every $a \in \mathbf{R}$, then $\phi = \alpha$ or ϕ is the identity mapping in the sense that *every* element of **C** is its own image. [Hint: Consider the possibilities for $\phi(i)$.]
4. Let S be an arbitrary ring and T the set of all ordered pairs (a, b) of elements of S. If addition and multiplication are defined by 3.43 and 3.44, respectively, verify each of the following.
   (i) T is a ring,
   (ii) T is a commutative ring if and only if S is a commutative ring,
   (iii) T has a unity if and only if S has a unity.

5. In the notation of the preceding exericse, let S be the ring $\mathbf{Z}$ of integers. Exhibit addition and multiplication rules for the corresponding ring T. Is T a field in this case? Is it an integral domain? Is it an ordered integral domain?

3.11 GEOMETRIC REPRESENTATION AND TRIGONOMETRIC FORM

It is implicit in our construction of the complex numbers that the mapping $a + bi \rightarrow (a, b)$ is a one-one mapping of the set $\mathbf{C}$ of all complex numbers onto the set of all ordered pairs of real numbers. Now in ordinary plane analytic geometry we represent points in the plane by their coordinates, that is, by ordered pairs of real numbers. Accordingly, we may represent a point in the plane by a single complex number. In other words, we shall sometimes find it convenient to associate with the complex number $a + bi$ the point with rectangular coordinates (a, b), and to say that this point has *coordinate* $a + bi$. A number of examples are given in Figure 8. It will be observed that a real number, that is, a complex number of the form $a + 0i$, is the coordinate of a point on the x-axis. A number of the form $0 + bi$, sometimes called a *pure imaginary*, is the coordinate of a point on the y-axis. We may also observe that a complex number $a + bi$ and its conjugate $a - bi$ are coordinates of points that are symmetrically located with respect to the x-axis.

Instead of specifying points in a plane by means of rectangular coordinates, we may of course use polar coordinates. If P is the point with nonzero coordinate $a + bi$, the distance of P from the origin O of coordinates is the positive real number $r = \sqrt{a^2 + b^2}$. If θ is an angle in standard position with terminal side OP, as in Figure 9, then by the definition of the trigonometric functions we have

$$a = r\cos\theta, \quad b = r\sin\theta.$$

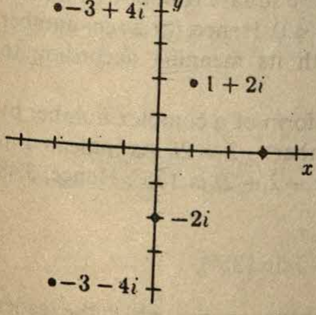


Figure 8

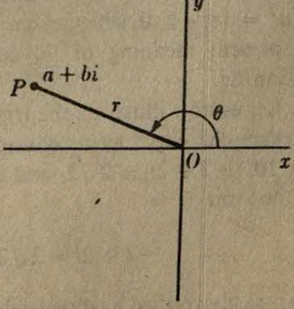


Figure 9

It follows that the complex number $a + bi$ can be expressed in the form

3.48 $$a + bi = r(\cos\theta + i\sin\theta).$$

We have been assuming that $a + bi \neq 0$. If $a + bi = 0$, then $r = 0$ in 3.48, and θ may be a completely arbitrary angle.

We now introduce some appropriate terms in the following definition.

3.49 Definition. The expression on the right side of 3.48 is called the ***trigonometric form*** of the complex number $a + bi$. The nonnegative real number $r = \sqrt{a^2 + b^2}$ is called the ***absolute value*** of the complex number $a + bi$ and may be designated by $|a + bi|$. The angle θ occurring in 3.48 is called ***an angle*** of $a + bi$.

Clearly, the nonnegative real number r occurring in the trigonometric form of $a + bi$ is uniquely determined. However, the angle θ is not unique, but if $r \neq 0$ and θ_1 and θ_2 are any two possible angles of $a + bi$, then elementary properties of the sine and cosine functions show that $\theta_1 = \theta_2 + n \cdot 360°$ for some integer n.

As a consequence of these observations, let us point out that if r and s are positive real numbers and we know that

$$r(\cos\theta + i\sin\theta) = s(\cos\phi + i\sin\phi),$$

then necessarily $r = s$ and $\theta = \phi + n \cdot 360°$ for some integer n.

We have previously defined (3.7) absolute values for an ordered integral domain, and we know that the field of complex numbers is not ordered. However, the present definition of absolute value is an extension of the concept for real numbers. For if a is a real number, we may consider it to be the complex number $a + 0i$ and, by 3.49, we have $|a| = \sqrt{a^2}$. But if c is a positive real number, by $\sqrt{c}$ we mean the positive square root of c. It follows that $\sqrt{a^2} = a$ if $a \geq 0$, whereas $\sqrt{a^2} = -a$ if $a < 0$. Hence, for a *real* number a, the present meaning of $|a|$ coincides with its meaning according to Definition 3.7.

Let us now illustrate the trigonometric form of a complex number by some examples. First, let us consider the number $-2 + 2i$. As indicated in Figure 10, $|-2 + 2i| = 2\sqrt{2}$, and an angle of $-2 + 2i$ is 135°. Hence, 3.48 takes the form

$$-2 + 2i = 2\sqrt{2}(\cos 135° + i\sin 135°),$$

which is easily verified by direct calculation. Other examples, which the reader

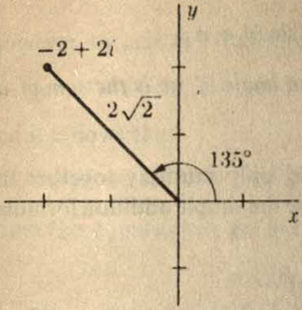


Figure 10

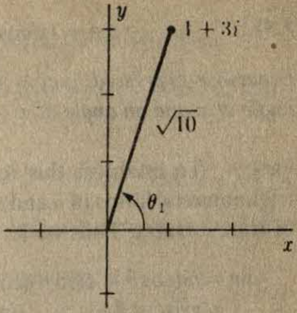


Figure 11

may check, are the following:

$$\begin{aligned} 1 + \sqrt{3}i &= 2(\cos 60^\circ + i \sin 60^\circ), \\ 4 &= 4(\cos 0^\circ + i \sin 0^\circ), \\ -i &= 1(\cos 270^\circ + i \sin 270^\circ), \\ -2(\cos 40^\circ + i \sin 40^\circ) &= 2(\cos 220^\circ + i \sin 220^\circ). \end{aligned}$$

It is only in special cases that we can find in degrees an angle of a given complex number. Naturally, an approximation may be obtained by use of trigonometric tables, or an angle may be merely indicated as in the following example. Let us attempt to express $1 + 3i$ in trigonometric form. Clearly, $|1 + 3i| = \sqrt{10}$, but we cannot exactly express its angle in degrees. However, if θ_1 is the positive acute angle such that $\tan \theta_1 = 3$, as indicated in Figure 11, we may write

$$1 + 3i = \sqrt{10}(\cos \theta_1 + i \sin \theta_1)$$

as the trigonometric form of $1 + 3i$.

The fact that every complex number can be expressed in trigonometric form is of considerable significance largely because of the following remarkable theorem.

3.50 Theorem. *If u and v are complex numbers such that in trigonometric form*

$$u = r(\cos \theta + i \sin \theta)$$

and

$$v = s(\cos \phi + i \sin \phi),$$

then the trigonometric form of uv is given by

3.51 $$uv = rs(\cos(\theta + \phi) + i\sin(\theta + \phi)).$$

Otherwise expressed, $|uv| = |u| \cdot |v|$, *and an angle of* uv *is the sum of an angle of* u *and an angle of* v.

PROOF. To establish this result we need only multiply together the trigonometric forms of u and v and then use the simple addition formulas of trigonometry. Thus we have

$$\begin{aligned} uv &= rs(\cos\theta + i\sin\theta)(\cos\phi + i\sin\phi) \\ &= rs[(\cos\theta\cos\phi - \sin\theta\sin\phi) + i(\cos\theta\sin\phi + \sin\theta\cos\phi)] \\ &= rs[\cos(\theta + \phi) + i\sin(\theta + \phi)], \end{aligned}$$

and the desired result follows immediately.

The special case of the preceding theorem in which $u = v$ shows at once that

$$u^2 = r^2(\cos 2\theta + i\sin 2\theta).$$

The following generalization of this result is of great importance.

3.52 De Moivre's Theorem. *If n is an arbitrary positive integer and*

$$u = r(\cos\theta + i\sin\theta),$$

then

3.53 $$u^n = r^n(\cos n\theta + i\sin n\theta).$$

PROOF. For $n = 3$, we use Theorem 3.50 and the case in which $n = 2$ which has just been established as follows:

$$\begin{aligned} u^3 = u^2 \cdot u &= [r^2(\cos 2\theta + i\sin 2\theta)][r(\cos\theta + i\sin\theta)] \\ &= r^3(\cos 3\theta + i\sin 3\theta). \end{aligned}$$

A general proof along these lines can be easily given by induction.

One application of this theorem will be given in the next section. However, let us point out here how certain trigonometric identities can be obtained in an easy way by use of this theorem. By letting $r = 1$ and, as an example, taking $n = 3$, we see that

$$(\cos\theta + i\sin\theta)^3 = \cos 3\theta + i\sin 3\theta.$$

However, by actually multiplying out the left side, we find that

$$(\cos\theta + i\sin\theta)^3 = \cos^3\theta - 3\cos\theta\sin^2\theta + i(3\cos^2\theta\sin\theta - \sin^3\theta),$$

and it follows that

$$\cos 3\theta + i\sin 3\theta = \cos^3\theta - 3\cos\theta\sin^2\theta + i(3\cos^2\theta\sin\theta - \sin^3\theta).$$

From this equation we get at once the two following trigonometric identities:

$$\cos 3\theta = \cos^3\theta - 3\cos\theta\sin^2\theta$$

and

$$\sin 3\theta = 3\cos^2\theta\sin\theta - \sin^3\theta.$$

EXERCISES

1. Express each of the following complex numbers in trigonometric form and indicate the points in a coordinate plane that have these numbers as coordinates:

   (a) $-1 - i$, (b) $-\sqrt{3} + i$,
   (c) $\sqrt{3} + i$, (d) $-1 + \sqrt{3}i$,
   (e) -4, (f) $3 - 2i$,
   (g) $2 - 2i$, (h) $\cos 26^\circ - i\sin 26^\circ$.

2. Express each of the following complex numbers in the form $a + bi$:

   (a) $4(\cos 45^\circ + i\sin 45^\circ)$,
   (b) $2(\cos 120^\circ + i\sin 120^\circ)$,
   (c) $3(\cos 180^\circ + i\sin 180^\circ)$,
   (d) $3(\cos 270^\circ + i\sin 270^\circ)$,
   (e) $\frac{1}{2}(\cos 300^\circ + i\sin 300^\circ)$,
   (f) $12(\cos 0^\circ + i\sin 0^\circ)$,
   (g) $11(\cos 90^\circ + i\sin 90^\circ)$,
   (h) $(\cos 117^\circ + i\sin 117^\circ)(\cos 123^\circ + i\sin 123^\circ)$.

3. Use De Moivre's Theorem to compute each of the following, and then express your answers in algebraic form by evaluating the necessary trigonometric functions:

   (a) $(-1 - i)^5$, (b) $(\sqrt{3} - i)^8$,
   (c) $(-i)^{12}$, (d) $\left(\frac{1}{\sqrt{2}} + \frac{i}{\sqrt{2}}\right)^{100}$
   (e) $\left(-\frac{1}{2} - \frac{\sqrt{3}i}{2}\right)^6$, (f) $(-1 + i)^{10}$,
   (g) $(1 - \sqrt{3}i)^{11}$, (h) $(\cos 18^\circ + i\sin 18^\circ)^{10}$.

4. Verify that the points with coordinates

$$(\cos 60^\circ + i\sin 60^\circ)^n \qquad (n = 1, 2, 3, 4, 5, 6)$$

   are the vertices of a regular hexagon inscribed in a circle of radius 1.

5. If u^* is the conjugate of the complex number u, verify each of the following:

   (a) $|u^*| = |u|$,
   (b) $uu^* = |u|^2$,
   (c) $u^{-1} = \dfrac{u^*}{|u|^2}$, if $u \neq 0$.

6. Show that if $u \neq 0$, De Moivre's Theorem also holds for every *negative* integer n.

7. Let $u, v \in \mathbf{C}$, and let P and Q be the points in a coordinate plane having respective coordinates u and v. Let R be the point with coordinate $u + v$. If O is the origin, show that OR is a diagonal of the parallelogram having OP and OQ as adjacent sides.

8. Show that if $u, v \in \mathbf{C}$, then $|u + v| \leq |u| + |v|$.

9. Use De Moivre's Theorem to find trigonometric identities for $\cos 4\theta$ and $\sin 4\theta$.

10. If $u = a + bi$, we have defined $|u| = \sqrt{a^2 + b^2}$. Use this definition to prove directly that if $u, v \in \mathbf{C}$, then $|uv| = |u| \cdot |v|$.

11. If $u, v \in \mathbf{C}$ with $v \neq 0$, prove that $\left|\dfrac{u}{v}\right| = \dfrac{|u|}{|v|}$.

3.12 THE *n*th ROOTS OF A COMPLEX NUMBER

In this section we give an important application of the use of the trigonometric form of a complex number. First, we give the following familiar definition.

3.54 Definition. Let n be a positive integer. If $u, v \in \mathbf{C}$ such that $v^n = u$, we say that v is an ***n*th root** of u.

We shall now prove the following theorem.

3.55 Theorem. *If n is a positive integer, and*

$$u = r(\cos\theta + i\sin\theta)$$

is a nonzero complex number in trigonometric form, there exist exactly n nth roots of u, namely,

$$\textbf{3.56}\qquad r^{1/n}\left(\cos\frac{\theta + k\cdot 360^\circ}{n} + i\sin\frac{\theta + k\cdot 360^\circ}{n}\right)\qquad (k = 0, 1, \ldots, n-1).$$

Here $r^{1/n}$ represents the principal nth root of the positive real number r, that is, the positive real nth root of r whose existence is asserted in Theorem 3.41.

PROOF. Suppose that $v = s(\cos\phi + i\sin\phi)$ is an nth root of u. Then $v^n = u$ and De Moivre's Theorem assures us that

$$s^n(\cos n\phi + i\sin n\phi) = r(\cos\theta + i\sin\theta).$$

It follows that the absolute values of the two members of this equation are equal, and an angle of one must be equal to an angle of the other. Hence $s^n = r$, so that $s = r^{1/n}$. Moreover, $n\phi = \theta + k\cdot 360°$ for some integer k, and it follows that $\phi = (\theta + k\cdot 360°)/n$. We have therefore shown that every nth root v of u must be of the form

$$v = r^{1/n}\left(\cos\frac{\theta + k\cdot 360°}{n} + i\sin\frac{\theta + k\cdot 360°}{n}\right) \tag{3.57}$$

for some integer k. Conversely, it is readily verified by De Moivre's Theorem that if v is given by 3.57, then $v^n = u$ for *every* choice of the integer k. The number of distinct nth roots of u is therefore the number of ways in which the integer k can be chosen in 3.57 so as to give distinct values of v. The angles obtained by letting k take the values $0, 1, \ldots, n-1$ have distinct terminal sides, and this fact makes it almost obvious that these n values of k yield distinct values of v. Moreover, if t is an arbitrary integer, the Division Algorithm (see Chap. 4) asserts that there exist integers q and r with $0 \le r < n$ such that $t = qn + r$, and therefore

$$\frac{\theta + t\cdot 360°}{n} = \frac{\theta + r\cdot 360°}{n} + q\cdot 360°.$$

It is then clear that the angle $(\theta + t\cdot 360°)/n$ has the same terminal side as the angle $(\theta + r\cdot 360°)/n$. Since $0 \le r < n$, we see that all possible different values of v are obtained if in 3.57 we let k take the values $0, 1, \ldots, n-1$. This completes the proof of the theorem.

As an example of the use of this theorem, let us find the fifth roots of the complex number $-2 + 2i$. First, we express this number in trigonometric form as follows:

$$-2 + 2i = 2^{3/2}(\cos 135° + i\sin 135°).$$

In the notation of the theorem, we have $r = 2^{3/2}$, $\theta = 135°$, and $n = 5$. Accordingly, the fifth roots of $-2 + 2i$ are the following:

$$\begin{gathered}
2^{3/10}(\cos 27° + i\sin 27°),\\
2^{3/10}(\cos 99° + i\sin 99°),\\
2^{3/10}(\cos 171° + i\sin 171°),\\
2^{3/10}(\cos 243° + i\sin 243°),\\
2^{3/10}(\cos 315° + i\sin 315°).
\end{gathered}$$

An interesting special case of Theorem 3.55 arises if we choose $u = 1$; hence $r = 1$ and $\theta = 0°$. We state this case as follows.

3.58 Corollary. *The distinct nth roots of* 1 *are the complex numbers*

3.59
$$\cos\frac{k\cdot 360°}{n} + i\sin\frac{k\cdot 360°}{n} \qquad (k = 0, 1, \ldots, n-1).$$

By De Moivre's Theorem, we have

$$\left(\cos\frac{360°}{n} + i\sin\frac{360°}{n}\right)^k = \cos\frac{k\cdot 360°}{n} + i\sin\frac{k\cdot 360°}{n}.$$

Hence the n distinct nth roots of 1, as given in 3.59, may all be expressed as powers of a certain nth root of 1. We have then the following alternate form of the preceding corollary.

3.60 Corollary. *Let us set*

3.61
$$w = \cos\frac{360°}{n} + i\sin\frac{360°}{n},$$

so that w is the nth root of 1 *having the smallest positive angle. Then the nth roots of* 1 *are the numbers*

3.62
$$w, w^2, w^3, \ldots, w^n = 1.$$

Since all nth roots of 1 have absolute value 1, they are coordinates of points on the circle with radius 1 and center the origin. Moreover, it is clear from 3.59 that they are the vertices of a regular polygon of n sides inscribed in this circle, with one vertex at the real number 1. This fact is of considerable importance in the study of the constructability of regular polygons with ruler and compass.

EXERCISES

1. Find the cube roots of 1 and express the answers in algebraic form. Draw a figure showing that these numbers are the coordinates of the vertices of a regular polygon of three sides (equilateral triangle).
2. Do the corresponding thing for the fourth roots of 1.
3. Do the corresponding thing for the eighth roots of 1.
4. Show that the sixth roots of 1 are the cube roots of 1 and their negatives.

5. Find the required roots and express the answers in algebraic form:

   (a) The cube roots of $-2 + 2i$.
   (b) The cube roots of $-8i$.
   (c) The fourth roots of -4.
   (d) The sixth roots of $-i$.
   (e) The fourth roots of $-1 - \sqrt{3}i$.
   (f) The square roots of $-1 + \sqrt{3}i$.

6. In each of the following, express the required roots in trigonometric form:
   (a) The fifth roots of 1.
   (b) The fourth roots of $-1 + i$.
   (c) The fourth roots of $\frac{1}{2} + \frac{\sqrt{3}i}{2}$.
   (d) The sixth roots of $1 - i$.
   (e) The square roots of $1 + 2i$.
   (f) The fourth roots of $16(\cos 12^\circ + i \sin 12^\circ)$.

7. Show that if v is any one of the *n*th roots of the nonzero complex number u, and w is given by 3.61, then $v, wv, w^2v, \ldots, w^{n-1}v$ are all the *n*th roots of u.

8. Show that the multiplicative inverse of an *n*th root of 1 is also an *n*th root of 1.

9. If $t \in C$ such that $t^n = 1$ but $t^m \neq 1$ for $0 < m < n$, t is called a *primitive* *n*th root of 1. Show each of the following:

   (i) The number w, defined in 3.61, is a primitive *n*th root of 1.
   (ii) If t is a primitive *n*th root of 1, then $1, t, t^2, \ldots, t^{n-1}$ are distinct and are all of the *n*th roots of 1.
   (iii) If t is a primitive *n*th root of 1, then t^l is also a primitive *n*th root of 1 if and only if l and n have no common factor greater than one.

Chapter 4

Factorization

This chapter begins with an introduction to polynomial rings with coefficients in a commutative ring. The main theme of the chapter is the study of factorization properties in rings of polynomials with coefficients in a field and in the ring of integers. Even though these rings appear to be superficially very different, the methods used to prove results about factorization of elements in the two rings have a common underlying idea—namely, the use of a division algorithm. We bring out this similarity by examining carefully the proofs of results for polynomial rings, extracting the essential ingredients, and then verifying that these ingredients are present in the ring of integers. Some other rings are considered also.

4.1 POLYNOMIAL RINGS

Let S be a commutative ring with unity. Heretofore we have used letters to denote sets or elements of sets, but we now use the letter x in a different way. It is not an element of S; rather, it is a symbol which we shall use in an entirely formal way. It is customary to call such a symbol an ***indeterminate.*** It is our purpose in this section to construct a ring which contains S and also has x as an element. This goal will motivate the definitions which we proceed to give.

Let x be an indeterminate and let us consider expressions of the form

4.1
$$a_0x^0 + a_1x^1 + a_2x^2 + \cdots + a_nx^n,$$

where n is some nonnegative integer and $a_i \in S$ $(i = 0, 1, \ldots, n)$. Such an expression is called "a *polynomial* in x with coefficients in S" or simply "a polynomial in x over S." If i is an integer such that $0 \leq i \leq n$, we say that a_i is the *coefficient* of x^i in the polynomial 4.1; also we say that a_ix^i is a *term* of the polynomial 4.1 with coefficient a_i.

At this stage we are to think of 4.1 as a purely formal expression. That is, the $+$ signs are not to be considered as representing addition in a ring, and neither is x^i to be considered as a product $x \cdot x \cdots x$ with i factors. Later on, after we have proved the existence of a ring which contains S as well as x, we shall see that in this larger ring we can make these familiar interpretations and thus justify the notation we are using. At the present time, we could logically use some such symbol as $(a_0, a_1, a_2, \ldots, a_n)$ to designate the polynomial 4.1, but the definitions of addition and multiplication of polynomials to be given below will seem more natural with the familiar notation used in 4.1.

In the exercises at the end of this section, we follow this alternate approach to the definition of the polynomial ring. It has the feature of defining the ring of polynomials over S without introducing any elements or symbols which are not already in S.

For the moment, let S be the ring $\mathbf{Z}$ of integers. Then the following are examples of polynomials in x over $\mathbf{Z}$:

(i) $2x^0 + (-3)x^1 + 4x^2$,
(ii) $3x^0$,
(iii) $0x^0 + 0x^1 + 4x^2$,
(iv) $0x^0 + 2x^1 + (-1)x^2 + 0x^3$.

In order to avoid writing so many terms with zero coefficients, we could agree in the third of these examples to write merely $4x^2$ with the understanding that x^0 and x^1 are assumed to have zero coefficients. Also, it would certainly agree with usual practice if we omitted the terms with zero coefficients in the fourth example and wrote $2x^1 + (-1)x^2$ to designate this polynomial. These simplifications will be possible under general agreements which we now make.

Let us designate the polynomial 4.1 over S by the symbol $f(x)$, and let $g(x)$ be the following polynomial over S:

4.2 $$b_0x^0 + b_1x^1 + \cdots + b_mx^m,$$

where $m \geq 0$ and $b_i \in S$ $(i = 0, 1, \ldots, m)$. By the *equality* of $f(x)$ and $g(x)$, written in the usual way as $f(x) = g(x)$, we shall mean that the expressions 4.1 and 4.2 are identical except for terms with zero coefficients. We therefore consider a polynomial as being unchanged by the insertion, or omission, of any number of terms with zero coefficients. In particular, with reference to the above examples, we may write

$$0x^0 + 0x^1 + 4x^2 = 4x^2$$

and

$$0x^0 + 2x^1 + (-1)x^2 + 0x^3 = 2x^1 + (-1)x^2.$$

Also, if we wish, we could write

$$3x^0 = 3x^0 + 0x^1 + 0x^2 + 0x^3,$$

and so on.

With this understanding about zero coefficients, if $f(x)$ is a polynomial over S and i is an *arbitrary* nonnegative integer, we may speak of the coefficient of x^i in $f(x)$. For example, in the polynomial $1x^0 + 2x^1 + 3x^2$ over Z, the coefficient of x^{10} is zero. This language often helps to simplify statements about polynomials. As an illustration, we may state again our definition of equality of two polynomials as follows. If $f(x)$ and $g(x)$ are polynomials over S, by $f(x) = g(x)$ we mean that for *every* nonnegative integer i, the coefficients of x^i in $f(x)$ and in $g(x)$ are equal elements of S.

Using the familiar sigma notation for sums, the polynomial 4.1 can be formally written as follows:

$$\sum_{i=0}^{n} a_i x^i.$$

Now let $S[x]$ denote the set of all polynomials in the indeterminate x over S. We proceed to define operations of addition and multiplication on the set $S[x]$. Of course, these definitions are suggested by the way that one adds and multiplies polynomials in elementary algebra. Let

4.3
$$f(x) = \sum_{i=0}^{n} a_i x^i$$

and

4.4
$$g(x) = \sum_{i=0}^{m} b_i x^i$$

be elements of $S[x]$. We define addition as follows: If r is the larger of m and n, then

4.5
$$f(x) + g(x) = \sum_{i=0}^{r} (a_i + b_i)x^i;$$

that is, for every nonnegative integer i, the coefficient of x^i in $f(x) + g(x)$ is the sum of the coefficients of x^i in $f(x)$ and in $g(x)$. Multiplication in $S[x]$ is defined as follows:

4.6
$$f(x)g(x) = \sum_{i=0}^{n+m} \left(\sum_{k=0}^{i} a_k b_{i-k} \right) x^i.$$

Another way of stating this definition of the product of $f(x)$ and $g(x)$ is to say that for each nonnegative integer i, the coefficient of x^i in the product is the sum (in the ring S) of all products of the form $a_r b_s$, where r and s are nonnegative integers such that $r + s = i$. The first few terms in the product given in 4.6 are as follows:

$$(a_0b_0)x^0 + (a_0b_1 + a_1b_0)x^1 + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots.$$

We are now ready to state the following theorem.

4.7 Theorem. *Let $S[x]$ be the set of all polynomials in the indeterminate x over the commutative ring S with unity. If operations of addition and multiplication are defined on $S[x]$ by 4.5 and 4.6, respectively, then*

(i) $S[x]$ is a commutative ring with unity,
(ii) $S[x]$ contains a subring isomorphic to S.

PROOF. The commutative and associative laws for addition in $S[x]$ follow from 4.5 since these laws hold in the ring S. Moreover, the polynomial $0x^0$ (which is equal to the polynomial with *all* coefficients zero) is the zero of $S[x]$ since, by 4.5, for each polynomial $f(x)$ we have

$$f(x) + 0x^0 = f(x).$$

Moreover, our definition of addition also shows that

$$\sum_{i=0} a_ix^i + \sum_{i=0} (-a_i)x^i = \sum_{i=0} [a_i + (-a_i)]x^i = 0x^0,$$

and each element of $S[x]$ has an additive inverse in $S[x]$.

To establish that multiplication is commutative, we observe that if $f(x)$ and $g(x)$ are given by 4.3 and 4.4, respectively, then the coefficient of x^i in $g(x)f(x)$ is

$$b_0a_i + b_1a_{i-1} + \cdots + b_ia_0,$$

and since S is assumed to be commutative, this is equal to the coefficient

$$a_0b_i + a_1b_{i-1} + \cdots + a_ib_0$$

of x^i in $f(x)g(x)$. Inasmuch as this statement is true for every nonnegative integer i, it follows that $f(x)g(x) = g(x)f(x)$, and hence that multiplication in $S[x]$ is commutative.

If $f(x)$ and $g(x)$ are given by 4.3 and 4.4, respectively, and

4.8 $$h(x) = c_0x^0 + c_1x^1 + \cdots + c_px^p$$

is also an element of $S[x]$, the coefficient of x^i in the product $(f(x)g(x))h(x)$ is found to be the sum of all products of the form $(a_rb_s)c_t$, where r, s, and t are nonnegative integers such that $r + s + t = i$. Similarly, the coefficient of x^i in the product $f(x)(g(x)h(x))$ is the sum of all products of the form $a_r(b_sc_t)$, with the same restriction on r, s, and t. However, since $(a_rb_s)c_t = a_r(b_sc_t)$ by the associative law of multiplication in S, it follows that

$$(f(x)g(x))h(x) = f(x)(g(x)h(x)),$$

that is, that multiplication is associative in $S[x]$.

We leave as exercises the proof of the distributive laws, and that if 1 is the unity of S, then $1x^0$ is the unity of $S[x]$. It follows then that $S[x]$ is a commutative ring with unity.

To establish part (ii) of the theorm, let S' denote the set of elements of $S[x]$ of the form ax^0, $a \in S$. It is easy to verify that S' is a subring of $S[x]$. Now the mapping $\theta: S' \to S$ defined by $\theta(ax^0) = a$, $a \in S$, is a one-one mapping of S' onto S. Moreover,

$$\theta(ax^0 + bx^0) = \theta[(a + b)x^0] = a + b = \theta(ax^0) + \theta(bx^0)$$

and

$$\theta[(ax^0)(bx^0)] = \theta[(ab)x^0] = ab = [\theta(ax^0)][\theta(bx^0)],$$

and it follows that θ is an isomorphism of S' onto S. Part (ii) of the theorem is therefore established.

Let us introduce some simplifications of our notation as follows. We shall henceforth identify S' with S, and therefore write simply a in place of ax^0; that is, we shall omit x^0 in writing polynomials. In particular, the zero polynomial will then be designated by the familiar symbol 0. We shall also write x in place of x^1, x^i in place of $1x^i$, and $-ax^i$ in place of $(-a)x^i$. We may now observe that x is itself an element of the ring $S[x]$. If $a \in S$, then also $a \in S[x]$, and ax^i can be interpreted as the product (in the ring $S[x]$) of a times x to the power i. Also, since each individual term of a polynomial 4.1 is itself equal to a polynomial, the + signs occurring in 4.1 can be correctly interpreted as addition in the ring $S[x]$. In other words, we have finally justified the use of the notation appearing in 4.1. Of course, addition is commutative in $S[x]$ and we can write the polynomial 4.1 with the terms in any order. For example, we could just as well write the polynomial 4.1 in the form

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

In this case it is customary to say that it is written in *decreasing* powers of x. As given in 4.1, it is written in *increasing* powers of x.

The following familiar concepts are of such great importance that we give a formal definition.

4.9 Definitions. The ***degree function deg*** is defined on the set of nonzero elements of $S[x]$ with values in $\mathbf{Z}^+ \cup \{0\}$, the set of non-negative integers, by the rule

$$\deg(s_0 + s_1 x + \cdots + s_n x^n) = n \quad \text{if} \quad s_n \neq 0.$$

Thus, $\deg(f(x)) = n$ if and only if $f(x) \neq 0$ and n is the highest power of x having a nonzero coefficient in $f(x)$. This nonzero coefficient of the highest power of x is called the ***leading coefficient*** of $f(x)$.

We emphasize that the zero polynomial has no degree.

The coefficient of x^0 in a polynomial is referred to as the ***constant term*** of the polynomial.

It will be observed that the nonzero elements of S, considered as elements of $S[x]$, are just the polynomials of degree zero.

If S is the ring $\mathbf{Z}$ of integers, the polynomials $2 + 3x - x^2$, $4x$, 3, and $x^4 - 2x$ have respective degrees 2, 1, 0, and 4 and respective leading coefficients -1, 4, 3, and 1. The constant terms are, respectively, 2, 0, 3, and 0.

We now show one of the ways the degree function is useful in the study of polynomial rings.

4.10 Theorem. *Let S be an integral domain. Then the polynomial ring $S[x]$ is an integral domain. Moreover, if $f(x)$ and $g(x)$ are nonzero elements of $S[x]$, then $f(x)g(x)$ is nonzero and*

$$\deg(f(x)g(x)) = \deg f(x) + \deg g(x). \tag{4.11}$$

PROOF. Since $f(x)$ and $g(x)$ are nonzero, they have degrees. Assume $\deg f(x) = n$ and $\deg g(x) = m$ and that the polynomials have the form

$$\begin{aligned} f(x) &= a_0 + a_1x + \cdots + a_nx^n, \\ g(x) &= b_0 + b_1x + \cdots + b_mx^m, \end{aligned}$$

with $a_n \neq 0$ and $b_m \neq 0$. It follows from the definition of multiplication that the term in the product involving the highest power of x is $a_nb_mx^{n+m}$. This term is nonzero because a_n and b_m are nonzero and S is an integral domain. It follows that the product $f(x)g(x)$ is nonzero and its degree is $n + m = \deg f(x) + \deg g(x)$. This proves both assertions in the theorem.

In this section we have introduced polynomials in *one* indeterminate x. However, this procedure can easily be generalized as follows. If S is a commutative ring with unity, then the polynomial ring $S[x]$ is a commutative ring with unity. If now y is another indeterminate, we may as above construct a ring $(S[x])[y]$ consisting of polynomials in y with coefficients in the ring $S[x]$. It is easy to verify that the elements of this new ring can also be expressed as polynomials in x with coefficients in the ring $S[y]$, in other words, that the rings $(S[x])[y]$ and $(S[y])[x]$ are identical. Accordingly, we may denote this ring by $S[x, y]$ and call its elements polynomials in the indeterminates x and y. Our last result, applied twice, shows that if S is an integral domain, then ring $S[x,y]$ is also an integral domain. These statements may be extended in the obvious way to polynomials in any number of indeterminates. However, we shall study mainly polynomials in just one indeterminate.

EXERCISES

1. Let S be a commutative ring with unity and let P denote the set of all sequences $(s_0, s_1, s_2, \ldots)$ in which each element $s_i \in S$, and, from some point on, all the s_i are zero; that is, for some nonnegative integer n (which may be different for different elements), $s_k = 0$ for all $k > n$. Define addition and multiplication on P by the rules

$$(s_0, \ldots, s_m, \ldots) + (t_0, \ldots, t_m, \ldots) = (s_0 + t_0, \ldots, s_m + t_m, \ldots)$$
$$(s_0, \ldots, s_m, \ldots) \cdot (t_0, \ldots, t_m, \ldots) = (u_0, \ldots, u_k, \ldots),$$

where the terms u_i are defined by

$$u_i = s_0 t_i + \cdots + s_j t_{i-j} + \cdots + s_i t_0.$$

Show that P is a commutative ring with unity.
Show moreover that the mapping $\theta: S[x] \to P$ defined by

$$\theta(a_0 + a_1 x + \cdots + a_n x^n) = (a_0, a_1, \ldots, a_n, 0, 0, \ldots)$$

is an isomorphism of $S[x]$ with P.

2. If S is a commutative ring with unity such that $S[x]$ is an integral domain, prove that S is an integral domain.

4.2 DIVISORS AND THE DIVISION ALGORITHM

In this section and in several more sections of this chapter we shall study the ring of polynomials with coefficients in an arbitrary *field* F. We know then, by Theorem 4.10, that $F[x]$ is an integral domain. Our goal is to study in some detail how elements of $F[x]$ are expressed as products. We begin with the following definition.

4.12 Definition. Let $F[x]$ be the ring of polynomials in the indeterminate x over an arbitrary field F. If $f(x), g(x) \in F[x]$, $g(x)$ is said to be a ***divisor*** (or ***factor***) of $f(x)$ if there exists $h(x) \in F[x]$ such that $f(x) = g(x)h(x)$. If $g(x)$ is a divisor of $f(x)$, we say also that $f(x)$ is ***divisible*** by $g(x)$ or that $f(x)$ is a ***multiple*** of $g(x)$ and indicate this by $g(x) \mid f(x)$ which is read as "$g(x)$ divides $f(x)$."

It follows immediately from this definition that if c is a nonzero element of F (that is, a polynomial of $F[x]$ of degree zero), then c is a divisor of every element $f(x)$ of $F[x]$. For, since c has a multiplicative inverse c^{-1} in F, we can write $f(x) = c(c^{-1}f(x))$, and this shows that c is a divisor of $f(x)$.

We shall sometimes write $g(x) \mid f(x)$ to indicate that $g(x)$ divides (that is, is a divisor of) $f(x)$.

It is important to observe that if $f(x) = g(x)h(x)$, then also $f(x) = (cg(x))(c^{-1}h(x))$, where c is any nonzero element of F. That is, if $g(x) \mid f(x)$, then $(cg(x)) \mid f(x)$ for every nonzero element c of F.

The next theorem states a property of $F[x]$ which will be used repeatedly to derive results about polynomial rings. The underlying idea is so useful that we shall isolate it and see how it can be applied in the study of other rings—the ring of integers, for example.

4.13 Division Algorithm. *If $f(x), g(x) \in F[x]$ with $g(x) \neq 0$, there exist unique elements $q(x)$ and $r(x)$ of $F[x]$ such that*

4.14 $$f(x) = q(x)g(x) + r(x), \quad r(x) = 0 \text{ or } \deg r(x) < \deg g(x).$$

We may recall that the zero polynomial has no degree and this fact explains the form of the condition which $r(x)$ is required to satisfy.

Before giving the proof of this theorem, we illustrate the statement by an example and indicate how the proof can be carried out. For the purpose of the example, let $F = Q$, the field of rational numbers, and let

$$f(x) = 2x^4 - x^3 + x^2 - x + 1, \quad g(x) = x^3 + 2x - 1.$$

If we set $q(x) = 2x - 1$ and $r(x) = 3x - 3x^2$, then

$$\begin{aligned} q(x)g(x) + r(x) &= (2x - 1)(x^3 + 2x - 1) + (3x - 3x^2) \\ &= 2x^4 - x^3 + x^2 - x + 1 \\ &= f(x), \end{aligned}$$

and $\deg r(x) = 2 < \deg g(x) = 3$.

This example simply verifies that the required $q(x)$ and $r(x)$ exist for the given $f(x)$ and $g(x)$. Let us indicate how one would actually carry out the computation starting with the polynomials $f(x)$ and $g(x)$.

Note first that $\deg g(x) < \deg f(x)$. We first find a multiple of $g(x)$ to make the highest degree term of $f(x)$ coincide with the highest degree term of the multiple of $g(x)$. In our case, $f(x)$ and $(2x)g(x)$ both have $2x^4$ as the highest degree term. Subtract these to obtain a polynomial of degree smaller than the degree of $f(x)$:

$$f_1(x) = f(x) - (2x)g(x) = -x^3 - 3x^2 + x + 1.$$

Since $f_1(x)$ has degree larger than the degree of $g(x)$, we repeat the first step: Find a multiple of $g(x)$ which has the same highest degree term as $f_1(x)$. In this case $(-1)g(x)$ and $f_1(x)$ have highest degree term $-x^3$, so we subtract and obtain a polynomial of smaller degree:

$$\begin{aligned} f_1(x) - (-1)g(x) &= -3x^2 + 3x \\ &= f(x) - (2x)g(x) - (-1)g(x) \\ &= f(x) - (2x - 1)g(x). \end{aligned}$$

The process stops at this point because the polynomial $r(x) = 3x - 3x^2$ has degree less than the degree of $g(x)$. The reader will notice that this process is nothing more than the usual method of long division of polynomials presented in a slightly different format. Let us formalize the precise idea which makes this procedure work.

4.15 Theorem. *Let $f(x)$ and $g(x)$ be nonzero elements of $F[x]$. If $\deg f(x) \geq \deg g(x)$, then there is an element $h(x)$ in $F[x]$ such that either $f(x) = g(x)h(x)$ or $\deg(f(x) - g(x)h(x)) < \deg f(x)$.*

Let us present a formal proof for this. Let the highest degree term of $f(x)$ be ax^n and the highest degree term of $g(x)$ be bx^m, with a and b nonzero elements of F. By assumption $\deg f(x) = n > m = \deg g(x)$. Thus $n - m$ is a nonnegative integer and x^{n-m} is an element of $F[x]$. For $h(x)$ we select $ab^{-1}x^{n-m}$. Then $f(x)$ and $g(x)h(x)$ have the same highest degree term. If all the terms are the same in both polynomials, then $f(x) = g(x)h(x)$; otherwise $f(x) - g(x)h(x)$ has degree less than $n = \deg f(x)$, as we were required to prove.

With this idea isolated, we turn to the proof of the Division Algorithm.

PROOF OF 4.13. Take the point of view that $g(x)$ is a fixed nonzero element of $F[x]$ and we must prove that for each element $f(x)$ in $F[x]$ the elements $q(x)$ and $r(x)$ exist in $F[x]$ with $f(x) = q(x)g(x) + r(x)$ and either $r(x) = 0$, or $\deg r(x) < \deg g(x)$.

There is an easy case we dispose of immediately. If $\deg f(x) < \deg g(x)$, then we may use $q(x) = 0$ and $r(x) = f(x)$.

If $\deg g(x) = 0$, then $g(x) = g$ is an element of F and has a multiplicative inverse. Thus, for any $f(x)$, we take $q(x) = g^{-1}f(x)$ and $r(x) = 0$. Henceforth, we assume $\deg g(x) \geq 1$.

For the rest of the proof we use mathematical induction. Let M be the set of all positive integers n with the property that the required $q(x)$ and $r(x)$ exist whenever $f(x)$ has degree n. We shall use 3.9 to show that $M = \mathbf{Z}^+$. Let us begin by showing that $1 \in M$. Take any $f(x)$ with $\deg f(x) = 1$. If $\deg g(x) \geq 2$, then this is the easy case already finished. If $\deg g(x) \leq 1$, then $\deg g(x) \leq \deg f(x)$ and so 4.15 may be applied. There is an element $h(x)$ with either $f(x) = g(x)h(x)$—in which case we take $q(x) = h(x)$, $r(x) = 0$—or

$\deg(f(x) - g(x)h(x)) < \deg f(x) = 1 = \deg g(x)$—in which case we again take $q(x) = h(x)$ and $r(x) = f(x) - g(x)h(x)$.

Now suppose M contains every positive integer less than or equal to c. It is necessary to show that $c + 1$ is in M. Let $f(x)$ be any element of $F[x]$ with $\deg f(x) = c + 1$. If $\deg g(x) > \deg f(x)$, then this case has already been done. Assume $\deg g(x) < c + 1$. We may apply 4.15 to obtain an element $h(x)$ such that either $f(x) = g(x)h(x)$—in which case we take $q(x) = h(x)$ and $r(x) = 0$—or $f_1(x) = f(x) - g(x)h(x)$ has degree less than $c + 1$. Thus $\deg f_1(x)$ is an element of M and so there exist elements $q_1(x)$ and $r_1(x)$ such that

$$f(x) - g(x)h(x) = f_1(x) = q_1(x)g(x) + r_1(x)$$

and either $r_1(x) = 0$ or $\deg r_1(x) < \deg g(x)$. We now observe that the choices $q(x) = q_1(x) + h(x)$, $r(x) = r_1(x)$ meet the required conditions. This shows that $c + 1$ is in M and it follows that $M = Z^+$. This proves the existence of the polynomials $q(x)$ and $r(x)$ for any given $f(x)$ and $g(x) \neq 0$. Now we must show the uniqueness of the q and r. Suppose a second pair of polynomials $q'(x)$ and $r'(x)$ meet the conditions of 4.14. Then we have the equations

$$f(x) = q(x)g(x) + r(x) = q'(x)g(x) + r'(x).$$

It follows that

$$(q(x) - q'(x))g(x) = r(x) - r'(x).$$

If this difference is not zero, then both sides of the equation have a degree. However, the left side of the equation has degree $\deg g(x) + \deg(q(x) - q'(x))$, which is at least as great as $\deg g(x)$. The right side of the equation has degree less than the degree of $g(x)$ because both $r(x)$ and $r'(x)$ have degree less than the degree of $g(x)$. This situation is impossible so $r(x) = r'(x)$ and then it also follows that $q(x) = q'(x)$.

It is customary to call $q(x)$ and $r(x)$ satisfying 4.14 the ***quotient*** and the ***remainder***, respectively, in the division of $f(x)$ by $g(x)$. Clearly, $f(x)$ is divisible by $g(x)$ if and only if the remainder in the division of $f(x)$ by $g(x)$ is zero.

EXERCISES

1. Find the quotient and remainder after division of $f(x)$ by $g(x)$ for the following choices of polynomials in $Q[x]$:

   (a) $f(x) = 4x^3 - x^2 - 3x + 5$, $g(x) = 2x + 5$;
   (b) $f(x) = x^5 - 2x^3 - x^2 - 3x - 6$, $g(x) = x^2 + 3$;
   (c) $f(x) = (x - 4)^3$, $g(x) = (x - 2)$;
   (d) $f(x) = (x - 2)^5$, $g(x) = (x - 1)^2$.

2. Let $g(x)$ be a nonzero element of $F[x]$ with $\deg g(x) \geq 1$. Show for each $f(x)$ in $F[x]$, there exist polynomials $r_0(x), r_1(x), \ldots$ which are either equal to zero or have degree

less than $\deg g(x)$ and which satisfy

$$f(x) = r_0(x) + r_1(x)g(x) + r_2(x)g(x)^2 + \cdots + r_n(x)g(x)^n$$

for some nonnegative integer n.

3. For each pair $f(x)$ and $g(x)$ listed in Exercise 1, find the polynomials $r_i(x)$ which satisfy the conditions in Exercise 2.
4. For the case $\deg g(x) = 1$, use the theory of Taylor series studied in calculus to find the polynomials $r_i(x)$ of Exercise 2 for any $f(x)$ in $Q[x]$.

4.3 THE DIVISION ALGORITHM IN OTHER RINGS

A careful examination of the statement and proof of the Division Algorithm for the polynomial ring $F[x]$ will show that there are two main ingredients: one is the integral domain, $F[x]$; and the second is the degree function defined on the set of nonzero elements of $F[x]$ and having nonnegative integers for its values. The property relating this function to the addition and multiplication in the ring is contained in the condition 4.15 and this in turn was the condition which led to the proof of the Division Algorithm. It will be seen as we progress through a series of results about polynomial rings and the ring of integers that the Division Algorithm plays a central role. Let us abstract this information from the previous situation and consider only these features.

4.16 Definition. Let R be an integral domain and let d be a function defined on the set of nonzero elements of R with values in the set $Z^+ \cup \{0\}$, the nonnegative integers. The function d is called a ***Euclidean function*** if the following condition is satisfied:

4.17 If f and g are elements of R and $g \neq 0$, then there exist elements $q, r \in R$ such that $f = gq + r$ and either $r = 0$ or $d(r) < d(g)$.

The elements q and r are called the *quotient* and *remainder* after division of f by g. We call a ring which possesses a Euclidean function a ***Euclidean domain.***

We shall examine some examples of integral domains and show them to be Euclidean domains. The following lemma provides one method for verifying that a function d is a Euclidean function. This is the analog of the method used for the case of the degree function for $F[x]$.

4.18 Lemma. *Let R be an integral domain and let d be a function defined on the set of nonzero elements of R and having values in the set of nonnegative integers. Suppose the following condition is satisfied:*

4.19 *For each pair of nonzero elements $f, g \in R$ for which $d(f) \geq d(g)$, there exists an $h \in R$ such that either $f = gh$ or $d(f - gh) < d(f)$. Then d is a Euclidean function and R is a Euclidean domain.*

PROOF. We must prove that the Division Algorithm property 4.17 is valid. Let $f, g \in R$ with $g \neq 0$. If $f = 0$, then 4.17 holds with $q = r = 0$. Assume that $f \neq 0$. If there is some h in R with $f = gh$, then $q = h$ and $r = 0$ is the right choice. Assume from now on that g does not divide f. In particular, no element $f - gx$ is zero. The set of nonnegative integers defined by

$$M = \{d(f - gx) : x \in R\}$$

must have a least element by the well ordering property. Select an element $t \in R$ such that $d(f - gt)$ is the least element of M. We set $r = f - gt$ and intend to show that $d(r) < d(g)$. Suppose this is not the case. Then $d(r) \geq d(g)$ and so we may apply the hypothesis 4.19 using r in place of f to conclude that there is some element $h \in R$ with either $r = gh$ or $d(r - gh) < d(r)$. We consider these two alternatives one at a time. Suppose $r = gh$. Then, substituting for r, $f - gt = gh$ and we conclude that $f = g(t + h)$; but this is not possible because we had assumed g was not a divisor of f. For the second alternative, again substitute for r to conclude

$$d(r - gh) = d(f - gt - gh) = d(f - g(t + h)) < d(r).$$

But the integer $d(f - g(t + h))$ is an element of M smaller than the smallest element of M, $d(r)$. This is impossible, so in fact $d(r) < d(g)$ as we wished to prove. Now rewrite the defining equation for r in the form $f = gt + r$, and we see the condition 4.17 holds with $q = t$ and this choice of r.

One may think of the function d as assigning "size" to each element of R. The condition 4.19 says that if f is "larger than" g, then either g divides f or one may subtract a suitable multiple of g from f to obtain an element of smaller "size" than f.

What has been gained by this general point of view? For one thing, we have isolated an ingredient which ensures the existence of a division algorithm. There are many other examples where these conditions are present. We shall give two examples.

For the first example we take the ring **Z** of integers. For the function d we take the absolute value; that is, d is defined by $d(x) = |x|$. It is quite easy to verify that **Z** is a Euclidean domain with d as its Euclidean function. To see this, we begin with two nonzero integers f and g for which $d(f) \geq d(g)$. Select h

to be either $+1$ or -1. The choice should ensure that hg and f have the same sign; that is, both are positive or both are negative. Then we have $d(f - gh) < d(f)$. Hence the condition 4.19 holds. Note that $d(x) = 0$ only if $x = 0$. Now we may conclude that there is a division algorithm for the ring of integers. Rather than giving an exact translation of 4.17 to this case, we give a slightly better form which conveys more information.

4.20 Division Algorithm for Integers. *Let f and g be integers with $g \neq 0$. Then there exist unique integers q and r such that $f = qg + r$ and $0 \leq r < |g|$.*

PROOF. Lemma 4.18 implies the existence of q_1 and r_1 such that $f = q_1g + r_1$ and $0 \leq |r_1| < |g|$. If it happens that this r_1 is nonnegative, then take $q = q_1$ and $r = r_1$. If r_1 is negative, then one of the integers $r_1 + g$ and $r_1 - g$ is positive because $|r_1| < |g|$. Let us suppose that $r_1 + g$ is positive. Then set $q = q_1 - 1$ and $r = r_1 + g$. It follows that $f = qg + r$ and $0 \leq r < |g|$. Similarly, if $r_1 - g$ is positive then $q = q_1 + 1$ and $r = r_1 - g$ give the right elements. The uniqueness assertion is left as an exercise for the reader.

The unique integers q and r which satisfy 4.20 are called, respectively, the *quotient* and the *remainder* in the division of f by g. It is important to observe that f is divisible by g if and only if the remainder in the division of f by g is zero.

For a second application of the Lemma 4.18, let G be the set of all complex numbers of the form $a + bi$, $a, b \in \mathbf{Z}$. Addition and multiplication are defined by the rules

$$\begin{aligned}(a + bi) + (c + di) &= (a + b) + (c + d)i,\\ (a + bi)\cdot(c + di) &= (ac - bd) + (ad + bd)i.\end{aligned}$$

Then G is a subring of the field of complex numbers and is surely an integral domain. This is called the ring of ***Gaussian integers.*** Next we define the function d on G by the rule $d(a + bi) = a^2 + b^2$. Thus all values of d are nonnegative integers and $d(x) = 0$ only when $x = 0$. We wish to show that G is a Euclidean domain and d is a Euclidean function. It is necessary to verify that the condition in Definition 4.16 is satisfied.

Let $f = a + bi$ and $g = c + di$ be two elements of G, $a, b, c, d \in Z$, and assume that $d(f) \geq d(g) \neq 0$. It is necessary to show that there is an element $h \in G$ with either $f = gh$ or $d(f - gh) < d(f)$. We will prove that h may be selected as one of the four elements ± 1, $\pm i$.

Suppose, on the contrary, that for each of the four choices of h, we have $d(f - gh) \geq d(f)$. Then we have four inequalities obtained by using the definition of d:

4.21
$$\begin{aligned}(a-c)^2+(b-d)^2 &\geq a^2+b^2 && (h=1),\\ (a+c)^2+(b+d)^2 &\geq a^2+b^2 && (h=-1),\\ (a+d)^2+(b-c)^2 &\geq a^2+b^2 && (h=i),\\ (a-d)^2+(b+c)^2 &\geq a^2+b^2 && (h=-i).\end{aligned}$$

One may expand the squared terms, cancel common terms from the two sides of the inequality, and rearrange the terms to find that these four inequalities are equivalent to the two inequalities

$$\begin{aligned}c^2+d^2 &\geq 2|ac+bd|,\\ c^2+d^2 &\geq 2|ad-bc|.\end{aligned}$$

Since the terms appearing on each side of the inequalities are nonnegative, we may square both sides and then add the resulting inequalities to obtain

$$2(c^2+d^2)^2 \geq 4[(ac+bd)^2+(ad-bc)^2].$$

When the right side is expanded, we find it is equal to

$$4(a^2+b^2)(c^2+d^2).$$

So finally we have the inequality

$$2(c^2+d^2)^2 \geq 4(a^2+b^2)(c^2+d^2).$$

This can hold only if

$$d(g)=c^2+d^2 \geq 2(a^2+b^2)=2d(f).$$

However, our original assumption required that $d(g) \leq d(f)$ and since $d(f) \neq 0$, this is an impossible situation. The impossibility was a result of the assumption that all four inequalities 4.21 were valid, and so at least one of the inequalities is false. This means for one of the choices of h, it holds that $d(f-gh) < d(f)$. This allows the possibility that $f=gh$. Since we have proved that G is a Euclidean domain with d as Euclidean function, the Division Algorithm must be true.

4.22 Division Algorithm for Gaussian Integers. *Let f and g be elements of G with $g \neq 0$. Then there exist q and r in G with $f=qg+r$ and $0 \leq d(r) < d(g)$.*

Example. Find the elements q and r in G for the pair $f=4+9i$, $g=2-4i$. In this case, we have $d(f)=4^2+9^2=97$, $d(g)=2^2+(-4)^2=20$. We test the four elements h to see if $d(f-gh)<97$. For $h=1$ this fails, but for $h=-1$ we find

$d(f+g) = 6^2 + 5^2 = 61$. Then set

$$f_1 = f + g = 6 + 5i.$$

Since $d(f_1) > d(g)$ we repeat this step with f_1 and g. Here we use $h = i$ because $d(f_1 - gi) = 2^2 + 3^2 = 22 < d(f_1)$. This is still not less than $d(g)$; so repeat once more using $f_2 = f_1 - gi$ and selecting $h = i$. Then

$$d(f_2 - ig) = (-2)^2 + 1^2 = 5.$$

Now work back through these equations:

$$\begin{aligned} r = 2 - i = f_2 - ig &= (f_1 - gi) - gi = f_1 - 2ig \\ &= (f + g) - 2ig = f + (1 - 2i)g. \end{aligned}$$

The required q is $-(1 - 2i)$.

There is a new feature provided by this Euclidean domain. The quotient and remainder that appear in the Division Algorithm are *not* unique. This can be seen by an easy example: $f = 3 + 5i$, $g = 2$. Then

$$\begin{aligned} f &= g(2 + 2i) + (-1 + i) \\ &= g(1 + 2i) + (1 + i). \end{aligned}$$

Since $d(-1 + i) = d(1 + i) = 2 < d(2) = 4$, there are two sets of candidates for the quotient and remainder: $(q, r) = (2 + 2i, -1 + i)$ or $(q, r) = (1 + 2i, 1 + i)$.

EXERCISES

1. For each pair $f, g \in Z$, find the quotient q, and remainder r, as given in the Division Algorithm for the Integers:

f	33	33	79	-79
g	9	-9	20	-20

2. For each pair $f, g \in G$, the ring of Gaussian integers, find a q and an r which are predicted to exist in the Division Algorithm for the Gaussian Integers:

f	$4 + 9i$	$4 + 9i$
g	3	$2 + i$

3. Let g be a fixed positive integer with $g > 1$. Prove that for any integer f there exist unique integers $r_0, r_1, \ldots$ which satisfy $0 \le r_i < g$ and

$$f = r_0 + r_1 g + r_2 g^2 + \cdots + r_n g^n,$$

for some nonnegative integer n. (The r_i are the "digits of f" in the base g. For example, if $g = 10$, and $f = 753$, then $r_0 = 3$, $r_1 = 5$, and $r_2 = 7$.)

4.4 GREATEST COMMON DIVISORS

Let F be any field and $F[x]$ the ring of polynomials with coefficients in F. In Section 4.2 we gave attention to divisors of $f(x)$ and the Division Algorithm. Now we consider the divisors of some $f(x)$ which are also divisors of a second polynomial $g(x)$.

4.23 Definition. Let $f(x)$ and $g(x)$ be elements of $F[x]$ with at least one of them nonzero. A ***common divisor*** of $f(x)$ and $g(x)$ is a polynomial $h(x)$ such that $h(x) \mid f(x)$ and $h(x) \mid g(x)$. A polynomial $d(x)$ is a **greatest common divisor** (g.c.d.) of $f(x)$ and $g(x)$ if

(i) $d(x)$ is a common divisor of $f(x)$ and $g(x)$.
(ii) Whenever $h(x)$ is a common divisor of $f(x)$ and $g(x)$, then $h(x) \mid d(x)$.

It is not obvious from the definition that a g.c.d. of $f(x)$ and $g(x)$ even exists. We shall show below that one does exist and also we shall see how to compute one. There will in fact be many g.c.d.s because of the following observation. If $h(x)$ divides $f(x)$ and if c is any nonzero element of F, then $ch(x)$ also divides $f(x)$; that is, $f(x) = h(x)k(x)$ implies $f(x) = ch(x) \cdot c^{-1}k(x)$. Thus if $d(x)$ is a g.c.d. of $f(x)$ and $g(x)$, then $cd(x)$ is one also. It is sometimes convenient to be able to refer to a unique member of this potentially large class of elements, so we introduce the following notion.

4.24 Definition. A nonzero element of $F[x]$ is called a ***monic polynomial*** if its leading coefficient is the unity element 1 of F.

Notice that if $h(x)$ is a nonzero polynomial with leading coefficient c, then $c^{-1}h(x)$ is a monic polynomial. If one insists that the g.c.d. of two polynomials be a monic polynomial, then it is unique, as will be seen shortly. If we occasionally refer to *the* g.c.d. of two polynomials, we mean the g.c.d. which is monic.

Now we prove the existence of the g.c.d. and give one of its important properties.

4.25 Theorem. *Let $f(x)$ and $g(x)$ be elements of $F[x]$ with $g(x) \neq 0$. Then there is a greatest common divisor of $f(x)$ and $g(x)$. Moreover, the g.c.d. is the nonzero polynomial of least degree which can be written as*

4.26 $$d(x) = f(x)u(x) + g(x)v(x)$$

for some elements $u(x)$ and $v(x)$ in $F[x]$.

PROOF. We define two finite sequences $\{f_i\}$ and $\{q_i\}$ in the following way. Let $f_1(x) = f(x)$ and $f_2(x) = g(x)$. The other terms are defined by the series of equations

$$
\begin{array}{ll}
(1) & f_1(x) = q_1(x)f_2(x) + f_3(x), \\
(2) & f_2(x) = q_2(x)f_3(x) + f_4(x), \\
 & \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
(n-2) & f_{n-2}(x) = q_{n-2}(x)f_{n-1}(x) + f_n(x), \\
(n-1) & f_{n-1}(x) = q_{n-1}(x)f_n(x).
\end{array}
$$

These equations are obtained by repeated application of the Division Algorithm; it is assumed that $f_2(x), \ldots, f_n(x)$ are nonzero,

$$\deg f_2(x) > \deg f_3(x) > \cdots > \deg f_n(x),$$

and that the remainder in Equation $(n-1)$ is zero. Surely such a point must eventually be reached because the degrees of the $f_i(x)$ form a strictly decreasing sequence of nonnegative integers; such a sequence cannot contain more than $\deg f_2(x)$ terms. Now we prove that this last nonzero remainder, $f_n(x)$, is a g.c.d. of the polynomials $f(x)$ and $g(x)$.

The proof uses the following idea. If u, v, w are elements of $F[x]$ with $u + v = w$, and if h is an element of $F[x]$ which divides two of these three elements, then h also divides the third element. For example, if $u = hu'$ and $w = hw'$, then $v = w - u = h(w' - u')$. Let us begin by showing that $f_n(x)$ is a common divisor of $f_1(x)$ and $f_2(x)$. From Equation $(n-1)$ we conclude that $f_n(x)$ divides $f_{n-1}(x)$. Thus $f_n(x)$ divides two of the three terms in Equation $(n-2)$. Hence we conclude that $f_n(x)$ also divides $f_{n-2}(x)$. Examine Equation $(n-3)$ in the same way to conclude $f_n(x)$ divides $f_{n-3}(x)$. Continue to use the equations until reaching Equations (2) and (1) to conclude $f_n(x)$ divides $f_2(x)$ and $f_1(x)$. Thus $f_n(x)$ is a common divisor of $f_1(x)$ and $f_2(x)$. Next it is necessary to show that $f_n(x)$ is a greatest common divisor. Let $h(x)$ be any polynomial which divides both $f_1(x)$ and $f_2(x)$. Then $h(x)$ divides two of the three terms in Equation (1), so $h(x)$ divides the third term, $f_3(x)$. Next use Equation (2) to conclude that $h(x)$ divides $f_4(x)$. Continue through the list to conclude that $h(x)$ divides each $f_i(x)$; in particular, $h(x)$ divides $f_n(x)$ and so $f_n(x)$ is a g.c.d. as required.

Now we set $d(x) = f_n(x)$ and show that $d(x)$ can be expressed in the form stated in the theorem. The method is to show that there exists two sequences of polynomials $\{u_i\}$ and $\{v_i\}$ for $3 \le i \le n$, such that

4.27

$$f_i(x) = f_1(x)u_i + f_2(x)v_i.$$

From Equation (1) we obtain

$$f_3(x) = f_1(x) - f_2(x)q_2(x)$$

so we set $u_3 = 1$, $v_3 = -q_2(x)$. Now assume that we have defined $u_3, \ldots, u_j$ and $v_3, \ldots v_j$ for $3 < j < n$. Then use Equation $(j+1)$ to obtain

$$\begin{aligned} f_{j+1}(x) &= f_{j-1}(x) - f_j(x)q_j(x) \\ &= (f_1(x)u_{j-1} + f_2(x)v_{j-1}) - (f_1(x)u_j + f_2(x)v_j)q_j(x) \\ &= f_1(x)(u_{j-1} - u_jq_j(x)) + f_2(x)(v_{j-1} - v_jq_j(x)). \end{aligned}$$

This indicates how to define the next terms:

$$\begin{aligned} u_{j+1} &= u_{j-1} - u_jq_j(x), \\ v_{j+1} &= v_{j-1} - v_jq_j(x), \end{aligned}$$

and it follows that 4.27 holds. The sequence stops at u_n, v_n which gives the required elements for the expression of $d(x)$.

The last statement to be proved is that about the "least degree." Suppose $m(x)$ is a nonzero polynomial in $F[x]$ and $m(x) = f_1(x)u + f_2(x)v$ for some u, v in $F[x]$. Since $d(x)$ is a g.c.d. of $f_1(x)$ and $f_2(x)$, it follows that $d(x)$ divides two of the three terms in this equation, and so $d(x)$ also divides $m(x)$. If $m(x) = d(x)h(x)$, then

$$\deg m(x) = \deg d(x) + \deg h(x) \geq \deg d(x).$$

So finally all parts of the theorem have been proved.

The sequence of steps which yields Equations (1) to $(n-1)$ is called the ***Euclidean Algorithm.*** It will be used again with the ring of integers.

Since the expression (4.26) will be used frequently, we shall say an element of the form $f(x)u + g(x)v$ is a *linear combination* of $f(x)$ and $g(x)$. Thus the theorem asserts that the g.c.d. of two polynomials can be expressed as a linear combination of the two polynomials.

Example 1: Let $f(x) = x^3 + 2x - 1$ and $g(x) = x^2 - 1$ be two elements of $\mathbf{Q}[x]$. Let us find the g.c.d. and express it as a linear combination of $f(x)$ and $g(x)$. The sequence from the Euclidean Algorithm is

$$\begin{aligned} f(x) &= g(x)(x+2) + (x+1), \\ g(x) &= (x+1)(x-1). \end{aligned}$$

The last nonzero remainder is $d(x) = x + 1$, and

$$d(x) = (x+1) = f(x) - g(x)(x+2)$$

gives the expression as a linear combination of $f(x)$ and $g(x)$.

Example 2: Let $f(x) = (x-2)^3$ and $g(x) = x - 1$ be elements of $\mathbf{Q}[x]$. By using the Taylor series expansion of $f(x)$ at the point 1, as studied in calculus, we find

$$\begin{aligned} f(x) &= f(1) + f'(1)(x-1) + (f''(1)/2)(x-1)^2 + (f'''(1)/6)(x-1)^3 \\ &= -1 + 3(x-1) - 3(x-1)^2 + (x-1)^3 \\ &= -1 + [3 - 3(x-1) + (x-1)^2]g(x). \end{aligned}$$

It follows that the constant polynomial -1 is a linear combination of $f(x)$ and $g(x)$ and, since no polynomial can have degree smaller than $0 = \deg(-1)$, the g.c.d. of $f(x)$ and $g(x)$ is 1 (the monic version of the polynomial -1).

In a numerical case, the actual calculations may often be simplified by the following observation. If $d(x)$ is the g.c.d. of $f(x)$ and $g(x)$, then also $d(x)$ is the g.c.d. of $af(x)$ and $bg(x)$, where a and b are nonzero elements of F. Hence, instead of Equation (1) in the proof of Theorem 4.25, we might use the similar equation obtained by dividing $af(x)$ by $bg(x)$. Similarly, instead of the second equation we might work with $dq(x)$ and $er(x)$, where d and e are nonzero elements of F; and so on for the other equations. This modification will not affect the validity of the arguments used to show that $f_n(x)$ is the g.c.d. of $f(x)$ and $g(x)$, and may greatly simplify the work involved. Let us give an illustration by finding the g.c.d. of the polynomials

$$f(x) = x^3 + \tfrac{1}{2}x^2 + \tfrac{1}{3}x + \tfrac{1}{6}$$

and

$$g(x) = x^2 - \tfrac{1}{2}x - \tfrac{1}{2}$$

over the field Q of rational numbers. In order to avoid fractions, we divide $6f(x)$ by $2g(x)$, obtaining

$$6f(x) = (3x + 3)[2g(x)] + 8x + 4,$$

so that $r(x) = 8x + 4$. If we now divide $2g(x)$ by $r(x)/4$, we see that

$$2g(x) = (x - 1)[r(x)/4].$$

Since $f_4(x) = 0$, the g.c.d. of $f(x)$ and $g(x)$ is obtained from the last nonzero remainder, namely $8x + 4$, by multiplying it by the multiplicative inverse of its leading coefficient. Hence the g.c.d. of $f(x)$ and $g(x)$ is $x + \frac{1}{2}$.

It is sometimes convenient to use the following terminology.

4.28 Definition. Two nonzero elements $f(x)$ and $g(x)$ of $F[x]$ are said to be ***relatively prime*** if their g.c.d. is 1.

EXERCISES

1. Find the g.c.d. of each of the following pairs of polynomials over the field Q of rational numbers, and express it as a linear combination of the two polynomials:
   (i) $2x^3 - 4x^2 + x - 2$ and $x^3 - x^2 - x - 2$,
   (ii) $x^4 + x^3 + x^2 + x + 1$ and $x^3 - 1$,
   (iii) $x^5 + x^4 + 2x^3 - x^2 - x - 2$ and $x^4 + 2x^3 + 5x^2 + 4x + 4$,
   (iv) $x^3 - 2x^2 + x + 4$ and $x^2 + x + 1$.

2. Find the g.c.d. of each of the following pairs of polynomials over the indicated field, and express it as a linear combination of the two polynomials:

   (i) $x^3 + (2i + 1)x^2 + ix + i + 1$ and $x^2 + (i - 1)x - 2i - 2$; field **C** of complex numbers,

   (ii) $x^2 + (1 - \sqrt{2})x - \sqrt{2}$ and $x^2 - 2$; field **R** of real numbers.

3. Let n be an integer with $n \geq 2$ and let $g_1(x), \ldots, g_n(x)$ be n elements of $F[x]$, not all equal to zero. Define the g.c.d. of these n polynomials to be a polynomial $h(x)$ which divides each $g_i(x)$ and is divisible by every polynomial in $F[x]$ which divides each $g_i(x)$. Use the results already proved for $n = 2$ and mathematical induction to show that the g.c.d. of $g_1(x), \ldots, g_n(x)$ exists and can be expressed as a linear combination $u_1 g_1(x) + \cdots + u_n g_n(x)$ for some elements u_i of $F[x]$.

4. Carry out the computation of the g.c.d. as defined in the previous exercise for the three polynomials listed and express it as a linear combination of these three polynomials;

$$\begin{aligned} g_1(x) &= x^3 - 7x - 6, \\ g_2(x) &= x^3 + 2x^2 - x - 2, \\ g_3(x) &= x^3 - 3x^2 - 4x + 12. \end{aligned}$$

5. Let $f(x)$ and $g(x)$ be nonzero elements of $F[x]$, where F is a field. If the field F' is an extension of the field F, then $F[x] \subseteq F'[x]$ and we may also consider $f(x)$ and $g(x)$ to be elements of $F'[x]$. Show that the quotient and the remainder in the division of $f(x)$ by $g(x)$ are the same whether these polynomials are considered as elements of $F[x]$ or of $F'[x]$. In particular, conclude that if there exists an element $h(x)$ of $F'[x]$ such that $f(x) = g(x)h(x)$, then $h(x) \in F[x]$.

6. Verify that the Division Algorithm (4.13) remains true if the field F is replaced by an arbitrary commutative ring S with unity, provided only that $g(x)$ is required to have as leading coefficient an element of S with a multiplicative inverse in S.

4.5 GREATEST COMMON DIVISORS IN EUCLIDEAN DOMAINS

In this section we examine the notion of greatest common divisors in Euclidean domains in general, and in the ring of integers, in particular.

Let R be a Euclidean domain (Definition 4.16) and d its Euclidean function. We give the following definitions which parallel those given in the case $R = F[x]$ and $d =$ the degree function.

4.29 Definitions. (1) Let $a, b \in R$ with $b \neq 0$. We say b is a ***divisor*** of a (or b is a ***factor*** of a) if there is an element $r \in R$ with $a = br$. We write $b \mid a$ to indicate that b is a divisor of a.

(2) An element $h \in R$ is a ***common divisor*** of the elements a and b of R if $h \mid a$ and $h \mid b$.

(3) An element d in R is a ***greatest common divisor*** (g.c.d.) of the elements a and b in R if

(i) $d \mid a$ and $d \mid b$, and
(ii) For any element h of R which is a common divisor of a and b, we have $h \mid d$.

The zero is excluded as a divisor for convenience and some additional ease in making statements later on. This is not a significant restriction since the zero element could only be a divisor of the zero element if the restriction $b \neq 0$ were removed from the definition.

Just as in the case of $R = F[x]$, one must prove that a g.c.d. of a and b exists. In addition, we must consider the possibility that more than one g.c.d. of a and b exists. The existence will be demonstrated shortly. In order to discuss the relation between the g.c.d.s of a and b, we introduce the following terminology.

4.30 Definition. An element u in a ring R with unity is called a ***unit*** if it has a multiplicative inverse; that is, u is a unit if and only if there is some v in R with $uv = vu = 1$ = the unity of R. The set of all units of R is denoted by $U(R)$.

For the polynomial ring $F[x]$, the units are the nonzero elements of the field F. For the ring Z of integers the only units are ± 1.

The next computation shows how units arise in the discussion of common divisors.

4.31 Theorem. *Let R be any integral domain and let $a, b \in R$.*

(i) If $a \mid b$, then au divides b for any element $u \in U(R)$.
(ii) If $a \mid b$ and $b \mid a$, then $a = bu$ for some $u \in U(R)$.

PROOF. If $a \mid b$, then $b = ar$ for some r in R. This implies $(au)(ru^{-1}) = b$ for any unit u; thus $au \mid b$ and (i) is true.

To prove (ii) suppose both $a \mid b$ and $b \mid a$. Then $b = ar$ and $a = bu$ for some elements $r, u \in R$. Combine these two equations to obtain $b = ar = bur$. The cancellation law may be applied since $b \neq 0$ to give $1 = ur$; it follows that both u and r are units of R, as we wished to prove.

We now indicate how the Euclidean Algorithm may be used in this general context to prove the existence of g.c.d.s and also obtain one of its important properties.

4.32 Theorem. *Let R be a Euclidean domain and d its Euclidean function. Let a, b be elements of R, not both equal to zero. Then there is a greatest common divisor of a and b in R. If g is one g.c.d. of a and b, then any other g.c.d. of a and b is equal to gu for some unit $u \in U(R)$. Any g.c.d., g, of a and b can be expressed as a linear combination $g = ar + bt$, for some $r, t \in R$.*

PROOF. Since at least one of a and b is nonzero, let us assume that $b \neq 0$. We shall define two sequences of elements a_i, q_i as follows. Begin with $a_1 = a$ and $a_2 = b$. Apply the Division Algorithm (4.17) to obtain a_3 and q_2 which satisfy $a_1 = q_2a_2 + a_3$, and either $a_3 = 0$ or $d(a_3) < d(a_2)$. If $a_3 \neq 0$, then apply the Division Algorithm again to obtain $a_2 = q_3a_3 + a_4$, with either $a_4 = 0$ or $d(a_4) < da_3)$. We repeat this until a zero remainder is reached and thereby obtain the series of equations

$$\begin{array}{lll} & a_1 = q_2a_2 + a_3, & d(a_3) < d(a_2), \\ & a_2 = q_3a_3 + a_4, & d(a_4) < d(a_3), \\ \textbf{4.33} & \quad \cdots\cdots & \quad \cdots\cdots \\ & a_{n-2} = q_{n-1}a_{n-1} + a_n, & d(a_n) < d(a_{n-1}), \\ & a_{n-1} = q_na_n. & \end{array}$$

Note that we are assured that a zero remainder is eventually reached because the sequence of values $d(a_i)$ is a strictly decreasing sequence of nonnegative integers; it cannot be infinitely long since there are only a finite number of integers between $d(a_2)$ and 0. We now ask the reader to use the argument from the proof of Theorem 4.25 to prove that this last nonzero remainder a_n is indeed a g.c.d of a and b and, moreover, that this g.c.d. can be expressed as a linear combination, $a_n = a_1u + a_2v$, for some $u, v \in R$.

Now that we know at least one g.c.d. of a and b exists, let us suppose that both g and g_1 are g.c.d.s of a and b. Then we apply the condition 3(ii) in Definition 4.29; g is a common divisor of a and b, so g divides every greatest common divisor; that is, $g \mid g_1$. Similarly, g_1 is a common divisor and must divide the g.c.d. g, $g_1 \mid g$. We then have the two conditions $g \mid g_1$ and $g_1 \mid g$. Now apply Theorem 4.31 to conclude that $g_1 = gu$ for some unit $u \in U(R)$.

It should be noted that once a particular g.c.d. has been expressed as a linear combination of a and b, $g = ar + bt$, then any other g.c.d. $g_1 = gu$ has the expression $g_1 = gu = aru + btu$ as a linear combination.

Now we may apply this general result to the ring $\mathbf{Z}$ of integers using the absolute value $d(x) = |x|$ as the Euclidean function. The set of units of $\mathbf{Z}$ consists of just the two elements ± 1. This means that if g is a g.c.d. of two integers a and b, the $-g$ is the only other g.c.d. of a and b. Exactly one of the elements $\pm g$ is positive, so in this case it is possible to define a unique greatest common divisor of a and b by always selecting the positive choice. It is customary to refer to the positive g.c.d. as *the* greatest common divisor.

Let us now give a numerical example. Suppose that we desire to compute the g.c.d. of the integers 26 and 382. By ordinary division we find that Equations 4.33 take the following form:

$$\begin{aligned} 382 &= 14\cdot 26 + 18, \\ 26 &= 1\cdot 18 + 8, \\ 18 &= 2\cdot 8 + 2, \\ 8 &= 4\cdot 2. \end{aligned}$$

In this case, 2 is the g.c.d. since it is the last nonzero remainder.

Not only is the Euclidean Algorithm useful in computing the g.c.d. of two integers, but it is also useful in expressing the g.c.d. of two integers as a linear combination of these integers. Let us now use the equations previously exhibited to express each of the remainders as a linear combination of 382 and 26. For simplicity, we shall write a in place of 382 and b in place of 26. The calculations are as follows:

$$\begin{aligned} 18 &= a - 14b, \\ 8 &= b - 18 = b - (a - 14b) = 15b - a, \\ 2 &= 18 - 2\cdot 8 = a - 14b - 2(15b - a) \\ &= 3a - 44b. \end{aligned}$$

Hence,

$$2 = 3(382) - 44(26),$$

and we have expressed the g.c.d. of 382 and 26 as a linear combination of these two integers.

We shall sometimes find it convenient to let (a, b) designate the g.c.d. of a and b. Thus, for example, we have that $(382, 26) = 2$. There can be no possible confusion with other uses of the number pair notation since the context will make it clear that we are considering the g.c.d. of two integers and not, for example, the coordinates of a point in the plane.

We shall frequently need to refer to a pair of integers with 1 as their g.c.d. Accordingly, we make the following definition.

4.34 Definition. The integers a and b are said to be ***relatively prime*** if and only if their g.c.d. is 1, that is, if and only if $(a, b) = 1$.

EXERCISES

Unless otherwise specified, the letters represent arbitrary nonzero integers.

1. Find the g.c.d. of each of the following pairs of integers and express it as a linear combination of the two integers: (i) 52 and 38, (ii) 81 and 110, (iii) 320 and 112, (iv) 7469 and 2387, (v) 10,672 and −4147.

2. Show that a and b are relatively prime if and only if 1 is expressible as a linear combination of a and b.
3. If $d = (a, b)$ and $a = a_1 d$, $b = b_1 d$, show that $(a_1, b_1) = 1$.
4. If m is a positive integer, show that $(ma, mb) = m(a, b)$.
5. If $x = yz + t$, prove that $(x, z) = (z, t)$.
6. Prove that $(a, bc) = 1$ if and only if $(a, b) = 1$ and $(a, c) = 1$.
7. Write out a formal proof that every remainder in Equations 4.33 is expressible as a linear combination of a and b . [Hint: Assume that this is false, and obtain a contradiction.]
8. If a, b, and n are given, prove that n is expressible as a linear combination of a and b if and only if $(a, b) \mid n$.
9. (i) Define the g.c.d. of *three* nonzero integers.
   (ii) Establish the existence of the g.c.d. of three integers by proving a result analogous to Theorem 4.32.
   (iii) If d is the g.c.d. of a, b, and c, show that $d = ((a, b), c) = ((a, c), b) = (a, (b, c))$.
   (iv) Let d be the g.c.d. of a, b, and c. If $a = a_1 d$, $b = b_1 d$, and $c = c_1 d$, show that 1 is the g.c.d. of the three integers a_1, b_1, and c_1.

4.6 UNIQUE FACTORIZATION IN $F[x]$

This section contains the main result about factorization of polynomials with coefficients in a field. Roughly speaking, it says that a polynomial may be factored as a product of certain polynomials which themselves cannot be factored further; in addition, there is only one way such a factorization can be done.

Throughout this section, F denotes an arbitrary field. We begin with a definition.

4.35 Definition. A polynomial $p(x)$ in $F[x]$ is *irreducible* if $\deg p(x)$ is positive and if $p(x)$ is not equal to the product of two elements of $F[x]$ each having positive degree.

Thus if $p(x)$ is an irreducible polynomial and if $p(x) = f(x)g(x)$, then either $\deg f(x) = 0$ or $\deg g(x) = 0$; this means either $f(x)$ or $g(x)$ is an element of F. If c is a nonzero element of the field F and $f(x) \in F[x]$, then we always have $f(x) = c^{-1}(cf(x))$, so that every polynomial of the form $cf(x)$ is a divisor of $f(x)$. It is easy to verify that a polynomial $f(x)$ of positive degree over F is an irreducible polynomial over F if and only if the *only* elements of $F[x]$ of positive degree that are divisors of $f(x)$ are of the form $cf(x)$, $c \neq 0$.

Since the degree of the product of two polynomials over F is the sum of the degrees of the factors, it follows at once from Definition 4.35 that *every element of $F[x]$ of the first degree is necessarily irreducible over F.*

We may emphasize that the possible divisors of $p(x)$ that are being considered in Definition 4.35 are those which are elements of $F[x]$; that is, they must have coefficients in F. For example, consider the polynomial $x^2 - 2$ over the field **Q** of rational numbers. Now $x^2 - 2$ cannot be factored into the product of two polynomials of the first degree in $\mathbf{Q}[x]$, and hence $x^2 - 2$ is an irreducible polynomial over **Q**. However, if we should consider the same polynomial as a polynomial over the field **R** of real numbers, we find that it is *not* irreducible over **R** since we have the factorization $x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2})$ with these factors of the first degree having coefficients in **R**. As this example shows, the concept of a polynomial being irreducible is relative to a specified field which contains the coefficients of the given polynomial.

In later sections we shall discuss irreducible polynomials over each of the familiar fields of elementary algebra.

The following lemma, which is called Euclid's Lemma when stated for the ring of integers in place of $F[x]$, is the key step in the proof of the uniqueness part of our main theorem. We prove it first in order not to interrupt the argument later.

4.36 Lemma. *If $f(x)$ and $g(x)$ are nonzero polynomials over the field F and $p(x)$ is an irreducible polynomial over F such that $p(x) \mid (f(x)g(x))$, then $p(x) \mid f(x)$ or $p(x) \mid g(x)$.*

PROOF. If $p(x) \mid f(x)$ then the conclusion holds and we are finished. Suppose then that $p(x)$ does not divide $f(x)$. Let $m(x)$ be the monic g.c.d. of $p(x)$ and $f(x)$. Since $m(x)$ divides the irreducible polynomial $p(x)$, we have either $m(x) = 1$, the constant polynomial, or $m(x) = cp(x)$ for some element $c \in F$. If the latter choice is correct, then, since $m(x) \mid f(x)$, we obtain $p(x) \mid f(x)$ in contradiction to our supposition. Thus $m(x) = 1$ and we may apply Theorem 4.25 to obtain the existence of elements $u, v \in F[x]$ such that

$$1 = uf(x) + vp(x).$$

Multiply this equation by $g(x)$ to obtain

4.37
$$g(x) = uf(x)g(x) + vp(x)g(x).$$

By assumption, $p(x)$ divides the product $f(x)g(x)$ and so $p(x)$ divides each term on the right side of 4.37. It follows that $p(x)$ divides $g(x)$ as we wished to prove.

If a polynomial $p(x)$ has the property that whenever $p(x)$ divides a product $f(x)g(x)$, then either $p(x)$ divides $f(x)$ or $p(x)$ divides $g(x)$, it is called a

prime polynomial. The lemma just proved shows that an irreducible polynomial is a prime polynomial. It is also easy to see that a prime polynomial must be irreducible. Suppose that $p(x)$ is prime and $p(x) = f(x)g(x)$. Then $p(x)$ divides the product of $f(x)$ and $g(x)$, so $p(x)$ must divide one of these. If $p(x)$ divides $f(x)$, then we have $\deg p(x) \leq \deg f(x)$ and so

$$\deg f(x) \geq \deg p(x) = \deg(f(x)g(x)) = \deg f(x) + \deg g(x).$$

This implies $\deg g(x) = 0$ and so $p(x)$ is irreducible. Since the terms *prime* and *irreducible* refer to the same polynomials, it is unnecessary to use two different terms. However, we shall have occasion to consider rings other than polynomial rings where the two concepts may not be the same. For now, we shall use the terms interchangeably when considering polynomials.

There is an easy extension of Lemma 4.36 to the case in which more than two factors are used.

4.38 Lemma. *If $p(x)$ is a prime polynomial which divides the product $f_1(x)\cdots f_k(x)$ of the k polynomials $f_1(x),\ldots,f_k(x)$, then $p(x)$ divides $f_i(x)$ for at least one $i \leq k$.*

PROOF. We give a proof using mathematical induction on the number of factors. The result is true for $k = 1$ in a trivial way; it is true for $k = 2$ by the case just completed. Assume the result is true when the number of factors is less than some integer k. Then we let $g(x)$ be the product of the $k - 1$ terms $f_2(x),\ldots,f_k(x)$. Our assumption is that $p(x)$ divides

$$f_1(x)g(x) = f_1(x)f_2(x)\cdots f_k(x).$$

By the Lemma 4.36, we have either $p(x) \mid f_1(x)$—in which case we are finished—or $p(x) \mid g(x)$. In this case $g(x)$ is the product of $k - 1$ terms and the lemma holds for $k - 1$ terms. Thus $p(x) \mid f_i(x)$ for some i with $2 \leq i \leq k$. By the principle of induction, the lemma is true for any number of factors.

Now we may state and prove the main result.

4.39 Theorem. *If $f(x)$ is a polynomial of positive degree over the field F and a is its leading coefficient, then there exist distinct monic prime polynomials $p_1(x),\ldots,p_k(x)$ $(k \geq 1)$ over F such that*

$$f(x) = a[p_1(x)]^{n_1}[p_2(x)]^{n_2}\cdots[p_k(x)]^{n_k}, \tag{4.40}$$

where the n's are positive integers. Moreover, such a factorization is unique except for the order of the factors.

PROOF. There are two main parts of the proof. The first is to show that the polynomial $f(x)$ has a factorization in the form 4.40; the second is to show that the uniqueness assertion is valid. The first step is carried out by mathematical induction. Let M be the set of positive integers n for which the following statement is true:

S_n: If $f(x)$ is a polynomial of degree n, then $f(x)$ has a factorization in the form 4.40 in which the $p_i(x)$ are monic prime polynomials.

We show $1 \in M$. If $\deg f(x) = 1$, then

$$f(x) = ax + b = a(x + ba^{-1})$$

and so S_1 is true since every polynomial $x + c$ is prime.

Next suppose that M contains $1, 2, \ldots, r-1$ for some integer $r \geq 2$. We must show $r \in M$. Let $f(x)$ be any polynomial of degree r. If $f(x)$ is a prime polynomial, then $p(x) = a^{-1}f(x)$ is a monic prime polynomial and $f(x) = ap(x)$ gives the correct factorization to verify that S_r is true for this $f(x)$. Suppose that $f(x)$ is not prime. Then $f(x) = u(x)v(x)$ for some elements in $F[x]$ with $\deg u(x) < r$ and $\deg v(x) < r$. Thus the existence of a factorization into a product of primes for $u(x)$ and $v(x)$ is assured by the validity of S_n for $n < r$. Hence there are monic prime polynomials $g_1(x), \ldots, g_t(x)$ and monic prime polynomials $h_1(x), \ldots, h_s(x)$ such that

$$u(x) = bg_1(x) \cdots g_t(x),$$
$$v(x) = ch_1(x) \cdots h_s(x)$$

with $b, c \in F$. It follows that

$$f(x) = u(x)v(x) = bcg_1(x) \cdots g_t(x)h_i(x) \cdots h_s(x).$$

After collecting equal primes, we see that $f(x)$ has a factorization of the form 4.40. Thus S_n is true for every n.

Now we turn to the uniqueness assertion. Suppose we have two sets of monic prime polynomials $p_1(x), \ldots, p_k(x)$ and $q_1(x), \ldots, q_m(x)$ (repetitions allowed) such that

4.41
$$p_1(x) \cdots p_k(x) = q_1(x) \cdots q_m(x).$$

Our plan is to show that each of the $p(x)$'s equals one of the $q(x)$'s. Begin by cancelling as much as possible; if $p_1(x)$ equals one of the $q_i(x)$, then both may be cancelled from the product. After pairing as many of the $p_i(x)$ with equal $q_j(x)$ as possible, we are left with a product of the same form as 4.41. In the most favorable case, every term on the left side has been matched with a term on the right side: Equation 4.41 now reads $1 = 1$, and there is nothing left to prove. In the unfavorable case, some terms are left on both the right and left sides of 4.41. We bring in Lemma 4.38. Since $p_1(x)$ divides the left side, it must divide the right side and the

Lemma 4.38 implies that $p_1(x)$ divides $q_i(x)$ for some i. Since $q_i(x)$ is a prime, it follows that $q_i(x) = cp_1(x)$ for some $c \in F$. However, we had assumed all these were monic polynomials, so $c = 1$. Thus $p_1(x) = q_i(x)$ and further cancellation may be done. It follows that this unfavorable case does not actually occur and each $p_i(x)$ is equal to some $q_j(x)$. Now to obtain exactly the statement in the theorem, we suppose $f(x)$ has a factorization 4.40 and another factorization

4.42
$$f(x) = a[q_1(x)]^{m_1} \cdots [q_t(x)]^{m_t}$$

with the $q_i(x)$ distinct, monic, prime polynomials. The coefficient a is the leading coefficient of $f(x)$ so it is the same in both 4.40 and 4.42. Now the right sides of Equations 4.40 and 4.42 are equal, and, after cancelling the constant term a, we have a product of monic prime polynomials as considered just above. Each $p_i(x)$ is equal to some $q_j(x)$ and the number of repetitions of $p_i(x)$ equals the number of repetitions of the corresponding $q_j(x)$. In other words, except for the order in which we write down the prime powers, the two factorizations are exactly the same.

We emphasize that the prime polynomials in 4.40 are restricted to be monic polynomials; hence the leading coefficient of the right side is just a, the leading coefficient of $f(x)$. This theorem has great theoretical significance as we shall see in the next chapter. It will not be so easy in practice to actually carry out the factorization of a given polynomial into the form 4.40. It may even be difficult to determine if a given polynomial is prime. We shall examine some special instances of this problem in later sections.

4.7 ROOTS OF POLYNOMIALS

In defining the polynomial ring $F[x]$, we have emphasized that the symbol x is not an element of the field, F. This differs from the point of view which is traditionally used in the study of functions in calculus, for example. In that view, an expression $f(X) = X^2 - 2X + 5$ is interpreted as a function defined for all real numbers x by the rule $f(x) = x^2 - 2x + 5$. Thus the value of the function f at the point x is the real number $f(x)$ and f is a function from the real numbers to the real numbers. Using this point of view, one may then discuss the problem of finding solutions to polynomial equations; find all real numbers x such that $f(x) = 0$. For certain polynomials, such as the $f(x)$ given above, there is no real number solution to this polynomial equation. However, if we allow x to take on complex values, then the equation has two solutions, $x = 1 \pm 2i$. This is a situation that frequently arises when attempting to solve polynomial equations. The factorization of polynomials plays a very important role in the discussion of this problem.

We wish to consider this situation from a fairly general point of view. In place of the real numbers and the complex numbers, we consider a field F and a larger field L containing F; in the terminology introduced in Section 3.5, F is a subfield of L and L is an extension field of F. We shall consider polynomials with coefficients in F and allow the indeterminant x to take on values from L. Let us make this notion precise.

Let L be an extension of the field F and let $c \in L$. Define the function θ_c from $F[x]$ to L by the rule:

$$\text{If } f(x) = a_0 + a_1x + \cdots + a_nx^n,$$

then

$$\begin{aligned}\theta_c(f(x)) &= a_0 + a_1c + \cdots + a_nc^n\\ &= f(c).\end{aligned}$$

We take this as the definition of $f(c)$ as well as the definition of θ_c. The map θ_c is called the *evaluation map at c*. According to the terminology introduced in Section 2.6, θ_c is a homomorphism of the ring $F[x]$ into the ring L. This is easy to verify; we shall give only an outline of the verification. Let $f(x)$ and $g(x)$ be elements of $F[x]$. Then

$$\begin{aligned}f(x) = \textstyle\sum a_ix^i,\quad g(x) &= \textstyle\sum b_jx^j,\\ \theta_c(f(x) + g(x)) &= \theta_c(\textstyle\sum a_ix^i + \sum b_jx^j)\\ &= \theta_c(\textstyle\sum (a_i + b_i)x^i)\\ &= \textstyle\sum (a_i + b_i)c^i\\ &= \textstyle\sum a_ic^i + \sum b_jc^j\\ &= \theta_c(f(x)) + \theta_c(g(x)).\end{aligned}$$

In an analogous way

$$\theta_c(f(x)g(x)) = f(c)g(c) = \theta_c(f(x))\theta_c(g(x)).$$

Thus, the properties required of a homomorphism hold for θ_c.

For a fixed element c of L, we shall denote by $F[c]$ the set of all elements of L which equal $\theta_c(f(x))$ $(= f(c))$ for some element $f(x)$ in $F[x]$. Thus $F[c]$ is the set of all elements of L which are equal to $a_0 + a_1c + \cdots + a_nc^n$ for some nonnegative integer n and some set of elements a_i in F. Because θ_c is a homomorphism, $F[c]$ is a subring of L. It need not be a subfield, however. In all cases, $F[c]$ is the smallest subring of L which contains F and c. It will be seen shortly that the subring $F[c]$ is a field if and only if c is the solution of some polynomial equation $g(x) = 0$ with $g(x) \in F[x]$; that is, c is a root of some polynomial in $F[x]$ in the sense of the following definition.

4.43 Definition. If $f(x) \in F[x]$ and c is an element of an extension field of F such that $f(c) = 0$, then c is called a ***root of*** $\boldsymbol{f(x)}$.

Consider the case $F = \mathbf{R}$, the field of real numbers, and $L = \mathbf{C}$, the field of complex numbers. Take $c = i$. Then $\mathbf{R}[i] = \mathbf{C}$ because every element of $\mathbf{C}$ can be expressed as $a + bi = \theta_i(a + bx)$. Of course, any element of $\mathbf{C}$ may equal $\theta_i(f(x))$ for many different choices of $f(x)$. For example,

$$\begin{aligned}\theta_i(2 + 3x) &= \theta_i(3 + 3x + x^2)\\ &= \theta_i(2 + 3x + (1 + x^2)h(x))\\ &= 2 + 3i.\end{aligned}$$

In this example, $h(x)$ may be any element of $\mathbf{R}[x]$ because

$$\theta_i((1 + x^2)h(x)) = (1 + i^2)h(i) = (1 - 1)h(i) = 0.$$

The crucial point is that i is the root of the polynomial $1 + x^2$, and so i is also a root of any polynomial $(1 + x^2)h(x)$. The polynomial $1 + x^2$ is the polynomial in $\mathbf{R}[x]$ of least degree having i as a root and this polynomial is irreducible in $\mathbf{R}[x]$. This example illustrates the statements of the next result.

4.44 Theorem. *Let L be an extension field of F and c an element of L which is the root of some nonzero polynomial in $F[x]$. Let $g(x)$ be a polynomial in $F[x]$ of least possible degree which has c as a root. Then*

(i) $g(x)$ is an irreducible polynomial in $F[x]$;
(ii) If $h(x)$ is a polynomial in $F[x]$ such that $h(c) = 0$, then $h(x) = g(x)q(x)$ for some $q(x) \in F[x]$.

PROOF. In order to show that $g(x)$ is irreducible, we suppose that $g(x) = u(x)v(x)$ and we must show that either $\deg u(x) = 0$ or $\deg v(x) = 0$. Evaluate at c to obtain $0 = g(c) = u(c)v(c)$. Since $u(c)$ and $v(c)$ are elements of the field L whose product is zero, one of the elements equals zero. To be specific, let us assume $u(c) = 0$. Then c is a root of $u(x)$; by assumption $g(x)$ has least degree of all polynomials having c as root, so $\deg g(x) \leq \deg u(x)$. On the other hand the equality $g(x) = u(x)v(x)$ implies $\deg g(x) = \deg u(x) + \deg v(x)$. Since the degree of a polynomial is never negative, we conclude $\deg g(x) \geq \deg u(x)$. It follows that $\deg g(x) = \deg u(x)$ and so $\deg v(x) = 0$ as we wished to prove.

Now suppose $h(x)$ is a nonzero polynomial and $h(c) = 0$. In order to show that $g(x)$ divides $h(x)$, we apply the division algorithm to obtain

$$h(x) = g(x)q(x) + r(x) \qquad r(x) = 0 \text{ or } \deg r(x) < \deg g(x).$$

Evaluate at c to obtain

$$h(c) = g(c)q(c) + r(c).$$

By assumption, $h(c) = g(c) = 0$ and so also $r(c) = 0$. If $r(x)$ is not the zero polynomial, then $r(x)$ is a polynomial of degree smaller than the degree of $g(x)$ and also having c as root. Since $g(x)$ was selected to have least possible degree of all polynomials having c as root, this case cannot occur. The alternative is that $r(x) = 0$ and so $h(x) = g(x)q(x)$ as required.

This result gives a description of all the polynomials in $F[x]$ which have c as a root; they are all multiples of some irreducible polynomial having c as a root. Let us apply this to the case in which c is an element of F rather than of some larger field containing F.

4.45 The Factor Theorem. Let c be an element of F. The elements of $F[x]$ which have c as a root are those of the form $(x - c)h(x)$ where $h(x)$ is any polynomial in $F[x]$.

PROOF. Since $c \in F$, $x - c$ is a polynomial in $F[x]$ which has c as a root and is clearly a polynomial of least degree with this property. Theorem 4.44 implies that every polynomial in $F[x]$ having c as a root has the form $(x - c)h(x)$ for some $h(x)$ and it is clear that every polynomial of this form has c as a root.

As an application of this result, we obtain a restriction upon the number of roots of a polynomial.

4.46 Theorem. *A polynomial in $F[x]$ of degree n has at most n roots in an extension field of F.*

PROOF. Let L be an extension field of F and let $c_1, \ldots, c_k$ be distinct elements of L which are roots of $f(x)$. We carry out the proof by mathematical induction on n, the degree of $f(x)$. If $n = 1$, then $f(x) = ax + b$ and this has only one root, namely $c_1 = -b/a$. Assume the theorem is true for polynomials over any field which have degree smaller than n, $n \geq 2$. Since $F \subseteq L$, we may regard $f(x)$ as an element of $L[x]$ and then apply the factor theorem. This yields $f(x) = (x - c_1)h(x)$ for some $h(x)$ in $L[x]$ and $\deg h(x) = n - 1$. For each index i, $2 \leq i \leq k$ we have

$$0 = f(c_i) = (c_i - c_1)h(c_i).$$

Since $c_i - c_1 \neq 0$ as the c_j were distinct elements of L, it follows that these $k - 1$ elements are all roots of $h(x)$. By the induction assumption, $h(x)$ has at most $n - 1$ roots, so $k - 1 \leq \deg h(x) = n - 1$. Thus $k \leq n$ as required.

The next corollary is a simple consequence of this.

4.47 Corollary. *Let $g(x)$ and $h(x)$ be polynomials in $F[x]$ and suppose that there is a set $\{c_1, \ldots, c_k\}$ of k distinct elements of F such that $g(c_j) = h(c_j)$ for each $j = 1, \ldots, k$. If $k > \deg g(x)$ and $k > \deg h(x)$, then $g(x) = h(x)$.*

PROOF. Let us set $f(x) = g(x) - h(x)$, and then we have $f(c_j) = 0$ for the k elements c_j. If $f(x) \neq 0$, then $f(x)$ has a degree which certainly cannot exceed the degrees of either $g(x)$ or $h(x)$; this would be a nonzero polynomial having more roots than its degree. This is not possible because of the last result, so $f(x) = 0$ and $g(x) = h(x)$, as required.

Let us consider an example of these ideas to show the relation between the roots of polynomials and the problem of factorization. We consider the field **Q** of rational numbers and the polynomial $f(x) \in \mathbf{Q}[x]$ given by

$$f(x) = x^3 - x^2 - 3x - 1.$$

If we want to express this polynomial as a product of irreducible polynomials as in Theorem 4.39, it is necessary to find the irreducible divisors of $f(x)$. By inspection, we find $f(-1) = 0$ and so by the factor theorem,

$$f(x) = (x + 1)h(x).$$

By long division we find $h(x)$ and obtain

4.48
$$f(x) = (x + 1)(x^2 - 2x - 1).$$

Now it is necessary to factor the quadratic term or determine that it is irreducible. By the quadratic formula we find the two roots are $1 \pm \sqrt{2}$. We shall give a formal proof later that $\sqrt{2}$ is not a rational number. For now we shall simply use this fact. Then $c = 1 + \sqrt{2}$ is a real number which is a root of $h(x) = x^2 - 2x - 1$. Since c cannot be the root of a polynomial of degree one in $\mathbf{Q}[x]$, it follows that $h(x)$ is a polynomial of least degree in $\mathbf{Q}[x]$ which has c as a root. By Theorem 4.44(i), it follows that $h(x)$ is an irreducible element of $\mathbf{Q}[x]$. Thus the factorization 4.47 of $f(x)$ is the one assured by Theorem 4.39.

Finally, we give the following application of the Euclidean Algorithm.

4.49 Theorem. *Let L be an extension field of F and let c be an element of L which is the root of some nonzero polynomial in $F[x]$. Then the ring $F[c]$ is a subfield of L.*

PROOF. Since $F[c]$ is already known to be a subring of L, all that remains to prove is that every nonzero element of $F[c]$ has a multiplicative inverse. Let z be a nonzero element of $F[c]$. There is a polynomial $u(x) \in F[x]$ such that $u(c) = z$. By assumption c is the root of some nonzero polynomial in $F[x]$, so by Theorem 4.44, there is an irreducible polynomial $g(x)$ in $F[x]$ with c as a root. Since $z = u(c) \neq 0$, $g(x)$ does not divide $u(x)$. Since $g(x)$ is prime, it follows that the g.c.d. of $g(x)$ and $u(x)$ is 1. By Theorem 4.25 there exist polynomials $a(x)$ and $b(x)$ in $F[x]$ such that $g(x)a(x) + u(x)b(x) = 1$. Now evaluate at c:

$$1 = g(c)a(c) + u(c)b(c) = 0 + zb(c).$$

This shows that $b(c)$ is the inverse of z and since $b(c) \in F[c]$, it follows that every nonzero element of $F[c]$ has a multiplicative inverse in $F[c]$; thus $F[c]$ is a field.

We may view this result as saying that certain prime polynomials in $F[x]$ have roots in extension fields of F. The deficiency here is that we may draw this conclusion only for those prime polynomials which arise by selecting the root c in advance. This does not convey much information. In the next chapter we shall prove a much better result; namely, if we start with any prime polynomial $g(x)$ in $F[x]$, there is some field extension of F in which $g(x) = 0$ has a solution. It will follow then that any polynomial equation $f(x) = 0$ has a solution, perhaps not in F, but in some extension field of F.

Example. Consider the real number $\alpha = 1 + \sqrt[3]{2}$. This number is the root of the polynomial $g(x) = (x - 1)^3 - 2$ in the ring $\mathbf{Q}[x]$ of polynomials over the rational numbers. We ask the reader to verify (or take on faith) the fact that α is not the root of any polynomial in $\mathbf{Q}[x]$ having degree 1 or 2. Thus $g(x)$ is a prime polynomial and the ring $\mathbf{Q}[\alpha]$ is a field. As an illustration of the proof of the last theorem we show how to express α^{-1} as $f(\alpha)$ for a suitable $f(x) \in \mathbf{Q}[x]$. In the role of $u(x)$ from the proof, we have $u(x) = x$. Then

$$g(x) + x(-x^2 + 3x - 3) = -3$$

and substituting α for x yields

$$0 + \alpha(-\alpha^2 + 3\alpha - 3) = -3$$

Thus

$$\alpha^{-1} = (\alpha^2 - 3\alpha + 3)/3.$$

As an important special case, one or more of the monic prime polynomials occurring in a factorization of $f(x)$ may be of the first degree. In particular, the Factor Theorem (4.45) assures us that $x - c$, $c \in F$, is a factor of $f(x)$ if and only if $f(c) = 0$, that is, if and only if c is a root of the polynomial $f(x)$. The following definition introduces a terminology which is sometimes convenient.

4.50 Definition. The element c of F is said to be a root of ***multiplicity*** $m \geq 1$ of the polynomial $f(x)$ over F if $f(x)$ is divisible by $(x - c)^m$ but not by $(x - c)^{m+1}$. A root of multiplicity two is called a ***double root***.

It follows that c is a root of $f(x)$ of multiplicity m if and only if in the factorization 4.40 of $f(x)$ one of the prime factors occurring is $x - c$ and, furthermore, it occurs with the exponent m.

EXERCISES

1. (a) Prove that a polynomial $f(x)$ of degree 2 or 3 over a field F is a prime polynomial over F if and only if the polynomial $f(x)$ has no root in F.
   (b) Show, by means of an example, that a corresponding statement does not hold for polynomials of degree 4.
2. Determine whether or not each of the following polynomials is prime over each of the given fields. If it is not prime, factor it into a product of prime factors over each given field. As usual, **Q** is the field of rational numbers, **R** the field of real numbers, and **C** the field of complex numbers.

   (a) $x^2 + x + 1$ over **Q**, **R**, and **C**;
   (b) $x^2 + 2x - 1$ over **Q**, **R**, and **C**;
   (c) $x^2 + 3x - 4$ over **Q**, **R**, and **C**;
   (d) $x^3 + 2$ over **Q**, **R**, and **C**;
   (e) $x^2 + \sqrt{15}$ over **R** and **C**;
   (f) $x^2 + 4$ over **Q** and **Q**$[i]$;
   (g) $x^2 + 2\sqrt{2}x + 2$ over **Q**$[\sqrt{2}]$ and **R**.
4. In each case the polynomial over the given field has as a root the specified element of the field. Find the multiplicity of this root and complete the factorization of the polynomial into prime factors over the given field.

   (a) $x^4 + x^3 - 3x^2 - 5x - 2$ over **Q**, root -1;
   (b) $x^4 + 2x^2 + 1$ over **C**, root i.
5. Find an irreducible polynomial in **Q**$[x]$ which has the number c as a root for each of the following choices of c:

   (a) $1 + \sqrt{2}$ (b) $3 - 4\sqrt{3}$ (c) $5 - 2i$ (d) $7 - \sqrt{5}$
6. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ be a polynomial over the field F and let $b \in F$. Denote by $f(x + b)$ the polynomial obtained from $f(x)$ by substituting $x + b$ for x;

   $$f(x + b) = a_0 + a_1(x + b) + \cdots + a_n(x + b)^n.$$

   Prove that $f(x)$ is irreducible if, and only if, $f(x + b)$ is irreducible. Moreover, show that if c is a root of $f(x)$ in some extension field of F, then $c - b$ is a root of $f(x + b)$.
7. If

   $$f(x) = \sum_{i=0} a_i x^i$$

is a polynomial over a commutative ring S, let us define the *derivative* $f'(x)$ of $f(x)$ as follows:

$$f'(x) = \sum_{i=1} i a_i x^{i-1}.$$

Prove that

$$[f(x) + g(x)]' = f'(x) + g'(x)$$

and that

$$[f(x)g(x)]' = f(x)g'(x) + f'(x)g(x).$$

8. Suppose that $f(x)$ is a polynomial over the field F and let $f'(x)$ be the derivative of $f(x)$ as defined in Exercise 7. Prove each of the following:
   (i) If an element c of F is a root of $f(x)$ of multiplicity greater than one, then c is also a root of the polynomial $f'(x)$.
   (ii) If an element c of F is a root of $f(x)$ of multiplicity one, then c is not a root of $f'(x)$.
   (iii) If $f(x)$ can be expressed as a product of elements of $F[x]$ of the first degree, then $f(x)$ and $f'(x)$ are relatively prime if and only if $f(x)$ has no root of multiplicity greater than one.

4.8 UNIQUE FACTORIZATION IN EUCLIDEAN DOMAINS

In this section we establish the Unique Factorization Theorem for Euclidean domains in general, and the ring of integers in particular.

4.51 Definition. Let R be any integral domain. An element $p \in R$ is called ***irreducible*** if in any factorization $p = ab$ with $a, b \in R$, either a or b is a unit of R.

In the case $R = F[x]$, the irreducible elements are the prime elements. In more general integral domains, the term *prime* is usually reserved for elements which have the property proved in Theorem 4.36 for the ring of polynomials.

4.52 Definition. Let R be any integral domain. An element $p \in R$ is called ***prime*** if p is not a unit in R, and whenever $a, b \in R$ have the property that $p \mid ab$, then necessarily either $p \mid a$ or $p \mid b$.

There do exist integral domains in which irreducible elements are not prime. An example will be found in the exercises following this section.

However, such an example cannot be a Euclidean domain in view of the following result.

4.53 Theorem. *If R is a Euclidean domain, then an element p of R is irreducible if and only if it is prime.*

PROOF. Assume p is an irreducible element of R and that $p \mid ab$, for $a, b \in R$. It is necessary to show that either $p \mid a$ or $p \mid b$. If $p \mid a$, then we are finished, so assume p does not divide a. Then 1 is a g.c.d. of a and p and so by Theorem 4.32, there exist $u, v \in R$ such that $1 = up + va$. Multiply this equation by b to obtain $b = bup + vab$. Since p divides ab, p divides the right side of this equation and so p divides the left side; that is, $p \mid b$, as required. We leave the converse statement as an exercise for the reader.

Yet one more point requires discussion before the Unique Factorization Theorem can be proved. An assertion about the uniqueness of factorization must take into account several problems which are illustrated by the following factorizations of the integer 10 in **Z**:

$$10 = 2 \cdot 5 = 5 \cdot 2 = (-2) \cdot (-5) = (-1) \cdot 2 \cdot (-5).$$

Of course, the first pair of factorizations differs only in the order in which the primes 2 and 5 occur; this is not a significant difficulty. However, in the third factorization some different primes appear and in the last factorization we find three terms instead of two. The reason for the existence of these different factorizations is the existence of the units ±1. The numbers 2 and -2 are both prime, but one is a unit multiple of the other.

In a more general integral domain R, if u is a unit of R and p is a prime element, then up is also a prime element. If p and q are two primes of R, then $pq = (up)(u^{-1}q)$ gives two factorizations of the element pq (assuming $u \neq 1$). In order to take this into account, it will be necessary to agree that p and up are regarded as sufficiently the same so that we will not give up the hope of a uniqueness statement. This difficulty was overcome in the case of $R = F[x]$ by insisting that we use only monic prime polynomials in the factorization theorem. This requirement has the effect of selecting one prime $p(x)$ from the many possibilities $up(x)$, for u a nonzero element of F. In the case $R = Z$, the ring of integers, we will again be able to make a uniform selection because $U(\mathbf{Z}) = \{\pm 1\}$. For any prime p, the set $\{p, -p\}$ contains only one positive element, so in the factorization statement below we shall use only positive primes. Unfortunately, it is not possible to make this kind of uniform choice in a general Euclidean domain. This will be taken into account by using the following notion.

4.54 Definition. Two elements a and b of the integral domain R are called ***associates*** if $a \mid b$ and $b \mid a$.

Note that Theorem 4.31 implies that a and b are associates if and only if there is a unit u in $U(R)$ such that $a = ub$.

For the case of primes in a Euclidean domain, associates arise when one of the primes divides the other.

4.55 Lemma. *If p and q are prime elements of the Euclidean domain R, then $p \mid q$ if and only if p and q are associates.*

PROOF. If $p \mid q$, then $q = pt$ for some $t \in R$. Since q is a prime and R is a Euclidean domain, q is irreducible; thus Definition 4.51 implies either p or t is a unit; p cannot be a unit because it is prime. Thus t is a unit and p is an associate of q. The converse is immediate from the definition.

Finally we can state our main result about factorization.

4.56 Unique Factorization Theorem for Euclidean Domains. *Let R be a Euclidean domain and f a nonzero nonunit of R. Then there exist prime elements $p_1, \ldots, p_k$ of R (not necessarily distinct) and a unit u such that*

4.57
$$f = up_1 \cdots p_k.$$

Moreover, if $q_1, \ldots, q_s$ are also primes of R and if v is a unit of R such that $f = vq_1 \cdots q_s$, then $k = s$ and the q_i may be renumbered so that for each i, p_i is an associate of q_i.

PROOF. There are two main steps in the proof: The first is to show that each element has a factorization into a product of irreducible elements times a unit; the second is to show the uniqueness of such a factorization. The first step is considerably more subtle than in the proof of Unique Factorization for $F[x]$. We begin by proving the following assertion:

4.58 If there is a nonzero, nonunit element $f \in R$ which is not itself irreducible and cannot be expressed as a product of irreducible elements of R, then there is an infinite sequence $f_1, f_2, f_3, \ldots$ of elements of R which satisfies

(i) $f = f_1$;
(ii) for each $i > 1$, $f_{i-1} = f_i h_i$ for some nonunit $h_i \in R$;
(iii) f_i is not a product of irreducible elements of R.

To construct this sequence begin with $f_1 = f$. Then f_1 is not irreducible and so it can be factored as a product $f_1 = ab$, with neither a nor b a unit. Now if it happens that both a and b can be expressed as a product of irreducible elements, then forming the product of these two expressions, we would obtain an expression of f_1 as a product of irreducible elements. Such a product does not exist, so either a or b is not a product of irreducible elements; to be specific let us suppose it is a. Then we set $f_2 = a$ and $h_2 = b$ and we have $f_1 = f_2h_2$, with h_2 not a unit. This construction may be repeated because f_2 is not a product of irreducible elements; $f_2 = f_3h_3$. This construction may be repeated to produce the unending sequence required by 4.58.

Now we use the fact that R is a Euclidean domain with Euclidean function d to show that no infinite sequence of elements of R can satisfy conditions 4.58. Let

$$M = \{d(xf_j) \mid x \in R, x \neq 0, j = 1, 2, 3, \ldots\}.$$

So M is a set of nonnegative integers and, by the well ordering property, M has a least element. Let $w = vf_k$ be an element of R such that $d(w)$ is the least integer in M; thus

4.59 $$d(w) = d(vf_k) \leq d(xf_j), x \in R, x \neq 0, j \geq 1.$$

Now use the Division Algorithm to divide each f_j by w;

$$f_j = wq_j + r_j, r_j = 0$$

or

$$d(r_j) < d(w).$$

We shall now prove that $r_j = 0$ for every j. Suppose this is not the case. Then there are two cases to consider. If $k \leq j$, then

$$f_k = f_{k+1}h_{k+1} = f_{k+2}h_{k+2}h_{k+1} = \cdots = f_jh_jh_{j-1}\cdots h_k.$$

Thus, more simply, $f_k = f_jx$, for some $x \in R$. Substitute this into the expression involving r_j to get

$$\begin{aligned} r_j &= f_j - wq_j = f_j - f_kvq_j = f_j - f_jxvq_j \\ &= f_j(1 - xvq_j). \end{aligned}$$

This says that r_j is a multiple of f_j and so $d(r_j)$ is in M. But $d(r_j)$ is less than $d(w)$, the smallest element of M. This situation is impossible. We examine the next possibility; assume $k > j$. By the same reasoning as just used, it follows that $yf_k = f_j$, for some $y \in R$. Once again we obtain

$$r_j = f_j - wq_j = yf_k - f_kvq_k = f_k(y - vq_k).$$

So r_j is a multiple of f_k and so $d(r_j)$ is smaller than the smallest element of M, an impossiblity. The only alternative to these situations is that every

$r_j = 0$. Now we exploit the fact that $r_{k+1} = 0$. We have two equations

$$f_k = f_{k+1}h_{k+1},$$
$$f_{k+1} = wq_{k+1} = f_k v q_{k+1}.$$

Combine these to get

$$f_k = f_k v q_{k+1} h_{k+1}$$

and so

$$1 = vq_{k+1}h_{k+1};$$

this implies that h_{k+1} is a unit, contrary to the condition 4.58(ii) stated above.

We recap this argument: If there is a nonzero, nonunit in R which is not a product of irreducible elements of R, then we get the infinite sequence f_i with certain properties. These properties cannot hold in the Euclidean domain; hence the infinite sequence does not exist, and there is no nonunit which cannot be written as a product of irreducible elements. This proves that for any nonzero nonunit $f \in R$, there is a factorization 4.57.

Now the uniqueness statement must be proved. Suppose the element f has the factorization 4.57 and also

4.60 $$f = vq_1 \cdots q_s, \qquad q_j \text{ a prime, } v \text{ a unit.}$$

Let us arrange the notation so that $k \leq s$; if this were not the case, then reverse the roles of the p_i and q_j. Then $p_1 \mid f$, so p_1 divides the right side of 4.60. Since p_1 is a prime, it must divide one of the factors; it cannot divide the unit v, so $p_1 \mid q_j$ for some j. Renumber if necessary to obtain $p_1 \mid q_1$. Thus p_1 and q_1 are associates by Lemma 4.55, and there is a unit v_1 with $p_1 = q_1 v_1$. After cancelling q_1, we obtain the equation

$$uv_1 p_2 \cdots p_k = vq_2 \cdots q_s.$$

Repeat this with p_2 and continue until all the p_j have been cancelled to obtain

$$uv_1 \cdots v_k = v \cdot (\text{product of remaining } q_j).$$

The number of uncancelled q_j is $s - k$, since each p_i was paired with one q_j. However, the term on the left is a unit, so any remaining terms on the right are also units. Since a prime cannot be a unit, there are no primes remaining on the right. Thus $k = s$, and each p_i is an associate of some q_j, as we were required to prove.

Let us now restate some of these general notions for the specific case of the ring of integers. A prime of Z is a nonzero nonunit which cannot be expressed as a product of two integers, neither of which is a unit—the same as an irreducible element in Z. The first few positive primes are

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, \ldots.$$

The theorem about unique factorization is so important in the case of the integers that it has become traditional to give it the following formidable title.

4.61 Fundamental Theorem of Arithmetic. *Every integer n which is not equal to* 0, ± 1 *can be written as a product $n = ep_1p_2\cdots p_k$ where $e = \pm 1$, and each p_i is a positive prime. When the integer n is given, the unit e and the primes p_i are uniquely determined except for the ordering of the p_i.*

Of course most of this theorem has been proved in the more general case of a Euclidean domain. One obtains the uniqueness in this case because if two prime integers, p and q, are associates, then $p = \pm q$. When we insist upon positive primes, then p and q are associate positive primes if and only if $p = q$. The sign $e = +1$ if $n > 0$, and $e = -1$ if $n < 0$.

Of course, the primes occurring in a factorization of an integer into prime factors need not all be distinct. By combining the equal primes, we see that every integer $a > 1$ can be expressed uniquely in the form

4.62 $$a = p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k},$$

where the p's are distinct positive primes and each of $n_1, n_2, \ldots, n_k$ is a positive integer. The right side of 4.62 may conveniently be called the *standard form* of the integer a. As an example, $2^2 \cdot 3 \cdot 5$ is the standard form of the integer 60.

4.9 SOME APPLICATIONS OF THE FUNDAMENTAL THEOREM

If a and c are positive integers and c is a divisor of a, then $a = cd$ for some positive integer d. If c and d are expressed as products of prime factors, then clearly a is a product of all prime factors of c times all prime factors of d. Moreover, the Fundamental Theorem then states that this gives the unique factorization of a as a product of prime factors. It follows that the only possible prime factors of c (or of d) are the primes that are factors of a. If then a is expressed in the standard form 4.62, any divisor c of a is necessarily of the form

$$c = p_1^{m_1}p_2^{m_2}\cdots p_k^{m_k},$$

where $0 \le m_i \le n_i$ $(i = 1, 2, \ldots, k)$. Conversely, any integer c of this form is clearly a divisor of a.

It is now easy to obtain the g.c.d. (a, b) of two integers a and b if both a and b are expressed in standard form. Clearly, (a, b) is the product of those primes which are factors of both a and b, each such prime occurring to the smaller of the two powers to which it occurs in a and in b. For example, $60 = 2^2 \cdot 3 \cdot 5$ and $252 = 2^2 \cdot 3^2 \cdot 7$. It follows that $(60, 252) = 2^2 \cdot 3$.

Previously we have had a method for computing the g.c.d. of two integers by the use of Euclid's algorithm; this method does not involve finding any prime factors of the given integers. From a computational point of view, the previous method may involve much less work than the present one since it may be exceedingly difficult to find the prime factors of fairly large numbers, and therefore difficult to express them in standard form. Exercise 1 at the end of this section is designed to illustrate this point.

We conclude this section by giving a formal proof that certain polynomial equations do not have solutions in the rational number field. The Fundamental Theorem of Arithmetic plays an important part in the proof.

4.63 Theorem. *Let n be an integer with $n > 1$ and let m be any positive integer. If the polynomial $x^n - m$ has a rational root, then that root is an integer and m is the nth power of an integer.*

PROOF. Suppose a/b is a rational number which satisfies $(a/b)^n = m$. We may assume that a and b are positive integers having g.c.d. equal to 1. If this were not the case, we could write a and b in standard form, then cancel any prime factors common to both without changing the fraction a/b. Then we have

4.64
$$a^n = mb^n.$$

Our goal is to show that $b = 1$. If $b \neq 1$; then b is divisible by some prime integer p. Then Equation 4.64 implies that p also divides a^n. By the Fundamental Theorem we reason that the prime divisors of a^n are exactly the same as the prime divisors of a; only the powers of the primes are different. Let us make this argument explicit: Suppose the standard form of a is

$$a = p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k}.$$

Then we have

$$a^n = p_1^{ne_1} p_2^{ne_2} \cdots p_k^{ne_k}.$$

By the uniqueness of the factorization, this must be the standard form of a^n. Thus p also divides a; this conflicts with our choice that g.c.d. of a and b equals 1. It follows that $b = 1$ and m is the nth power of the integer a.

Here are two applications of this result. The real number $\sqrt{2}$ is not rational. The number $\sqrt{2}$ is the solution of the equation $x^2 - 2 = 0$. The theorem says that any rational solution must be an integer and 2 must be the square of an integer. However, 2 is a prime and the unique factorization of integers tells us that 2 is not the square of any integer. Hence $\sqrt{2}$ is not rational.

The number $1 + \sqrt[3]{2}$ is not rational. If it were rational, then adding the rational number -1 to it would produce the rational number $\sqrt[3]{2}$. By the theorem, this is rational only if 2 is the third power of an integer. Again since 2 is prime, this is not the case.

More general results along these lines will be considered when we discuss rational roots of polynomial equations having rational coefficients.

EXERCISES

1. Express each integer a, b listed below in standard form and then use this to find the g.c.d. of a and b. Compare the amount of calculation required with the amount of calculation required to find the g.c.d. using the Euclidean Algorithm:

a	120	970	53,599
b	4851	3201	133,331

2. (i) Using a method similar to that used in the proof of Lemma 4.53, show that if a is a divisor of bc and $(a, b) = 1$, then a is a divisor of c.
   (ii) Prove the same result by use of the Fundamental Theorem.
3. Show that a positive integer $a > 1$ is a perfect square (that is, is the square of an integer) if and only if in the standard form of a all the exponents are even integers.
4. Show that if b and c are positive integers such that bc is a perfect square and $(b, c) = 1$, then both b and c are perfect squares.
5. If n is a positive integer and $p_1, p_2, \ldots, p_n$ are distinct positive primes, show that the integer $(p_1 p_2 \cdots p_n) + 1$ is divisible by none of these primes.
6. Use the previous exercise to prove there exist infinitely many prime integers. Hint: Consider the possibility that $p_1, \ldots, p_n$ is the set of all primes.
7. List the 25 positive prime integers less than 100. For the ambitious reader, we mention that there are 95 positive primes less than 500. For the reader with a computer, we mention that there are 669 positive primes less than 5000.
8. Let R be any integral domain. Prove that a prime element of R is an irreducible element of R.
9. Let $R = \{a + b\sqrt{-5} : a, b \in Z\}$. Then R is a subring of the field $Q[\sqrt{-5}]$ and so R is an integral domain. Prove that R contains an irreducible element which is not prime. Do this by showing first that the integer 3 is an irreducible element of R.

Then show 3 is not prime because 3 divides

$$21 = (1 + 2\sqrt{-5})(1 - 2\sqrt{-5})$$

but 3 does not divide either factor on the right side.

10. Let G be the ring of Gaussian integers, $G = \{a + bi : a, b \in \mathbf{Z}\}$. Then G is a Euclidean domain (Section 4.3) and so the results of this section apply to G. Show $1 + i$ is a prime in G. Show that the integer 2 is not prime in G (even though it is a prime of $\mathbf{Z}$) and that $1 + i$ is one of the prime factors of 2. Show that the prime integer 3 is also a prime of G.

4.10 RATIONAL ROOTS OF A POLYNOMIAL OVER THE RATIONAL FIELD

If $f(x)$ is a polynomial of degree $n > 0$ over a field F, then clearly $f(x)$ and $cf(x)$ have the same roots for any nonzero element c of F. If, in particular, $f(x)$ has coefficients in the field $\mathbf{Q}$ of rational numbers and we choose c as any integer divisible by each of the denominators of the coefficients of $f(x)$, then $cf(x)$ will have coefficients that are integers. In studying the roots of a polynomial with rational coefficients there is therefore no loss of generality in restricting attention to polynomials that have integral coefficients. We shall now prove the following theorem.

4.65 Theorem. *Let*

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0, \qquad a_n \neq 0,$$

be a polynomial of positive degree n with coefficients that are integers. If r/s is a rational number, in lowest terms, which is a root of the polynomial $f(x)$, then r is a divisor of a_0 and s is a divisor of a_n.

We may recall that by saying that r/s is in lowest terms we mean that r and s are relatively prime integers and $s > 0$. However, the requirement that s be positive plays no role in the proof of this theorem.

PROOF. Since r/s is assumed to be a root of $f(x)$, we have that

$$a_n\left(\frac{r}{s}\right)^n + a_{n-1}\left(\frac{r}{s}\right)^{n-1} + \cdots + a_0 = 0.$$

If we multiply throughout by the nonzero integer s^n, we obtain

$$a_n r^n + a_{n-1} r^{n-1} s + \cdots + a_1 r s^{n-1} + a_0 s^n = 0. \tag{4.66}$$

By transposing the last term to the right side, this equation can be written in the form

$$(a_n r^{n-1} + a_{n-1} r^{n-2} s + \cdots + a_1 s^{n-1})r = -a_0 s^n.$$

Since all letters here represent integers, we see that the integer $a_0 s^n$ is divisible by the integer r. But we are given that r and s are relatively prime, and it therefore follows that a_0 is divisible by r.

By a similar argument, if in 4.66, we transpose $a_n r^n$ to the other side, we can see that a_n is divisible by s.

As an example of the use of this theorem, let us find all rational roots of the polynomial

$$g(x) = 4x^5 + x^3 + x^2 - 3x + 1.$$

If r/s is a rational number, in lowest terms, which is a root of this polynomial, then r must be a divisor of 1 and s a positive divisor of 4. It follows that $r = \pm 1$, $s = 1$, 2, or 4; and we see that the only possible rational roots are the following: $1, \frac{1}{2}, \frac{1}{4}, -1, -\frac{1}{2}, -\frac{1}{4}$. It is easy to verify by direct calculation that $g(1) \neq 0$, $g(\frac{1}{2}) = 0$, $g(\frac{1}{4}) \neq 0$, $g(-1) = 0$, $g(-\frac{1}{2}) \neq 0$, and $g(-\frac{1}{4}) \neq 0$. Hence, $\frac{1}{2}$ and -1 are the *only* rational roots. If we divide $g(x)$ by $x - \frac{1}{2}$ and then divide the quotient by $x + 1$, we find that

$$g(x) = (x - \tfrac{1}{2})(x + 1)(4x^3 - 2x^2 + 4x - 2).$$

Any root of this third degree factor is naturally a root of $g(x)$, so its only possible rational roots are therefore $\frac{1}{2}$ and -1. It is easy to verify that $\frac{1}{2}$ is a root and if we again divide by $x - \frac{1}{2}$, we can express $g(x)$ in the form

$$g(x) = (x - \tfrac{1}{2})^2(x + 1)(4x^2 + 4)$$

or

$$g(x) = 4(x - \tfrac{1}{2})^2(x + 1)(x^2 + 1). \tag{4.67}$$

We see therefore that $\frac{1}{2}$ is a double root of $g(x)$. Since the quadratic polynomial $x^2 + 1$ has no rational root, it is a prime polynomial over **Q** and hence in 4.67 we have $g(x)$ expressed as a product of prime polynomials over **Q**. For that matter, the polynomial $x^2 + 1$ is prime over the field **R** of real numbers and so 4.67 also gives the factorization of $g(x)$ into prime polynomials over **R**.

EXERCISES

1. Complete the proof of Theorem 4.65 by showing that s is a divisor of a_n.
2. Prove the following corollary of Theorem 4.65. A rational root of a *monic* polynomial with coefficients that are integers is necessarily an integer which is a divisor of the constant term of the polynomial.

3. Find the factorization of the polynomial $g(x)$ of the example given above into prime factors over the field **C** of complex numbers.

4. Find all rational roots of each of the following polynomials over the rational field **Q**:

   (a) $3x^3 + 5x^2 + 5x + 2$,
   (b) $2x^4 - 11x^3 + 17x^2 - 11x + 15$,
   (c) $x^5 - x^4 - x^3 - x^2 - x - 2$,
   (d) $x^3 + x^2 - 2x - 3$,
   (e) $6x^3 - 7x^2 - 35x + 6$,
   (f) $x^5 + 5x^4 + 13x^3 + 19x^2 + 18x + 8$,
   (g) $x^3 - (\frac{1}{5})x^2 - 4x + \frac{4}{5}$,
   (h) $x^7 + x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$.

5. Find all rational roots of each of the following polynomials over the rational field **Q**, and factor each polynomial into a product of prime polynomials over **Q**:

   (a) $9x^4 + 6x^3 + 19x^2 + 12x + 2$,
   (b) $x^5 - x^4 - 3x^3 + 6x^2 - 4x + 1$,
   (c) $4x^4 + 20x^3 + 33x^2 + 20x + 4$,
   (d) $2x^4 + 3x^3 + 4x + 6$.

6. Show that each of the following polynomials over **Q** has no rational root:

   (a) $x^{1000} - x^{500} + x^{100} + x + 1$,
   (b) $x^{12} - x^9 + x^6 - x^3 + 1$,
   (c) $x^m + 2x^{m-1} - 2$, m a positive integer ≥ 2.

*4.11. INTEGER SOLUTIONS OF LINEAR EQUATIONS

The solution of an equation $ax = b$, with $a, b \in \mathbf{Z}$ is given by $x = b/a$. This yields a rational solution so long as $a \neq 0$. If we ask for an integer solution, then, since the equation has only one rational solution, either there is no integer solution (in the case a does not divide b) or the unique rational solution happens to be an integer. The equation $3x = 7$ has no integer solution; the equation $3x = 6$ has a unique rational solution which happens to be an integer: $x = 2$. Suppose we consider a similar equation having two unknowns in place of one. We ask for the integer solutions of the equation $ax + by = c$, where a, b, c are given integers and the unknowns are x and y. This will have rational solutions so long as not both a and b are zero. The problem we intend to solve is that of deciding if there are integer solutions, and if there are, to find all of them. It may not be immediately obvious what the situation is; for example, do there exist integers x and y such that

4.68
$$1547x + 3367y = 455?$$

Let us consider a less dramatic example. The equation $2x + 6y = 19$ cannot have any integer solutions because no matter what integer values are substituted for x and y, the left side of the equation is an even integer (that is, it is divisible by 2) whereas the right side is not an even integer. This simple idea turns out to be the key to this problem.

We shall give a theorem which shows how to systematically determine if a solution exists and how to find all solutions when they do exist.

4.69 Theorem. *Let a and b be nonzero integers and g = the g.c.d. of a and b.*

(*1*) *For an integer c, the equation $ax + by = c$ has a solution with $x, y \in \mathbf{Z}$ if and only if $g \mid c$.*

(*2*) *Suppose $c = gc_1$, $a = ga_1$, $b = gb_1$ with $c_1, a_1, b_1 \in \mathbf{Z}$. Then every solution to the equation $ax + by = c$ with $x, y \in \mathbf{Z}$ is obtained by the following method: let x_1 and y_1 be integers which satisfy $a_1x_1 + b_1y_1 = 1$. Then*

4.70
$$\begin{aligned} x &= x_1c_1 - b_1t, \\ y &= y_1c_1 + a_1t, \end{aligned}$$

where t is any integer gives every solution to $ax + by = c$.

PROOF. If x and y are integers such that $ax + by = c$, then g = the g.c.d. of a and b must divide c since it divides each term on the left side of the equation. Thus, when a solution exists, g divides c. On the other hand suppose g divides c so that $c = gc_1$, for some $c_1 \in \mathbf{Z}$. There exist integers u and v such that $au + bv = g$, by Theorem 4.32. Now multiply by c_1 and use $x = uc_1$, $y = vc_1$ to get $ax + by = c$. Thus when g divides c, integer solutions to the equation exist. This proves statement (1).

Before we begin the proof of statement (2), it should be noted that the numbers a_1 and b_1 are relatively prime (see Exercise 3 following Section 4.5) and so there do exist integers x_1 and y_1 such that

4.71
$$a_1x_1 + b_1y_1 = 1.$$

If we set $x = x_1c_1 - b_1t$, $y = y_1c_1 + a_1t$, for any integer t, then

$$\begin{aligned} ax + by &= ga_1(x_1c_1 - b_1t) + gb_1(y_1c_1 + a_1t) \\ &= g(a_1x_1 + b_1y_1)c_1 - gb_1a_1t + gb_1a_1t \\ &= g(1)c_1 \\ &= c. \end{aligned}$$

This choice of x and y gives integer solutions to the equation. Next it is necessary to show that every solution is obtained this way. Let x and y be any integers such that $ax + by = c$. Since a, b, c are each divisible by g, we may cancel this factor to obtain the equation

4.72
$$a_1x + b_1y = c_1.$$

Next we multiply Equation 4.71 by c_1 to obtain

$$a_1x_1c_1 + b_1y_1c_1 = c_1. \tag{4.73}$$

Subtract Equation 4.73 from 4.72 to obtain

$$a_1(x - x_1c_1) + b_1(y - y_1c_1) = 0. \tag{4.74}$$

From this equation we conclude that a_1 divides $b_1(y - y_1c_1)$. However a_1 and b_1 have no common factor greater than 1, so a_1 divides $(y - y_1c_1)$; let us write $ta_1 = y - y_1c_1$, for some integer t. Substitute this into Equation 4.74 to obtain

$$a_1(x - x_1c_1) + b_1(a_1t) = 0.$$

It follows that $x - x_1c_1 = -b_1t$. After we rearrange some terms in these last equations, the form of the solution 4.70 is easily seen to be true.

Now let us apply this theorm to determine if Equation 4.68 does have a solution. First apply the Euclidean Algorithm to find the g.c.d. of 1547 and 3367:

$$\begin{aligned} 3367 &= 1547\cdot 2 + 273, \\ 1547 &= 273\cdot 5 + 182, \\ 273 &= 182\cdot 1 + 91, \\ 182 &= 91\cdot 2. \end{aligned}$$

So 91 is the last nonzero remainder, $91 = (3367, 1547)$. By working back through these equations, one finds

$$91 = 3367\cdot 6 + 1547\cdot(-13).$$

Since 455 is divisible by 91, Equation 4.68 has integer solutions and they are all given by

$$\begin{aligned} x &= 5(-13) - 37t = -65 - 37t, \\ y &= 5\cdot 6 + 17t = 30 + 17t, \qquad t \in Z. \end{aligned}$$

With the choice $t = -2$, we have $x = 9$, $y = -4$ as one solution of Equation 4.68.

EXERCISES

1. Find all integer solutions to the equations listed:

   (i) $4x + 7y = 19$,
   (ii) $14x + 49y = 70$,
   (iii) $144x + 256y = 10000$.

2. State and prove a theorem which gives a necessary and sufficient condition for the equation $ax + by + cz = d$ to have integer solutions for the unknowns x, y, z and given integers a, b, c, d.

3. The monetary system of a rather unusual country is such that all items for purchase are priced in a unit called the cento. For some unexplained reason, the country is able to produce only two types of coins; a 4 cento piece and a 7 cento piece. What are the possible number of centos that a purchaser can pay for an item exactly without the need for change? (What numbers N can be expressed as $4x + 7y = N$ with x and y nonnegative?)

Chapter 5

Field Extensions

In this chapter we introduce a fundamental algebraic construction which describes, in some sense, all the homomorphic images of a given ring. This construction will be used to solve several problems. By applying it to the ring $F[x]$ of polynomials over a field, it will be possible to show that every polynomial equation $f(x) = 0$ has a solution—not necessarily in F, but in some field containing F. By applying it to the ring of integers, we construct a large class of finite rings which appear in many algebraic problems. Some of these rings will be finite fields. When these two results are combined, we shall be able to describe all finite fields and study some of their properties which are not shared by all fields.

5.1 SPLITTING FIELDS

Let F be a field and $f(x)$ an element of positive degree in $F[x]$. We want to consider the polynomial equation $f(x) = 0$ and the question of the existence of solutions of this equation. We have seen examples where the equation has no solution in the field F but a solution did exist in a larger field L containing F. The classical example occurs with $F = \mathbf{R}$, $L = \mathbf{C}$, and $f(x) = x^2 + 1$. Our goal is to show that given any polynomial equation $f(x) = 0$, there is always some extension field L in which the equation has a solution. In order to motivate the construction of L, let us consider the situation in which L is known to exist and there is an element $c \in L$ with $f(c) = 0$. We saw in Section 4.7 that there is an irreducible polynomial $g(x)$ in $F[x]$ with $g(c) = 0$ and $g(x)$ dividing $f(x)$. Moreover, the subring $F[c]$ of L is actually a field; it consists of all elements of the form

$$z = a_0 + a_1c + a_2c^2 + \cdots + a_{n-1}c^{n-1} \tag{5.1}$$

where $n = \deg g(x)$ and all the a_i are elements of F. Moreover, given the element z, the expression 5.1 for z in this form is unique. Since $F[c]$ is a field that obviously contains the solution c of $f(x) = 0$, we may restrict our attention to the field $F[c]$. Our method will be to show how the field $F[c]$ may be constructed without knowing in advance that the field L containing a solution c exists. The evaluation map at c will be used to indicate the main ideas. Recall from Section 4.7 that the map $\theta_c : F[x] \to L$ is a homomorphism defined by $\theta_c(h(x)) = h(c)$ and $F[c]$ is the set of all elements of the form

$$z = \theta_c(h(x)) \tag{5.2}$$

for some $h(x)$ in $F[x]$. This provides us with this first important observation:

When the field $F[c]$ containing a solution of $g(x) = 0$ exists, this field is a homomorphic image of the polynomial ring $F[x]$.

Next we examine this more closely to see how a suitable homomorphic image of $F[x]$ is constructed. Every element z in $F[c]$ has the form 5.1 and also the form 5.2. For a given z, however, there are many choices for the polynomial $h(x)$. If $h(x)$ is one choice, then $h(x) + g(x)$ is another. This is so because

$$\begin{aligned} 0 &= g(c) = \theta_c(g(x)), \\ z &= \theta_c(h(x)) + 0 = \theta_c(h(x) + g(x)). \end{aligned}$$

By using the same reasoning, we find that additional choices may be made by taking the elements $h(x) + g(x)a(x)$, for any polynomial $a(x) \in F[x]$. Let us now verify that these are the *only* choices. Let $b(x)$ be any polynomial in $F[x]$ for which $z = \theta_c(b(x)) = \theta_c(h(x))$. Then by the homomorphism property of θ_c we conclude

$$\theta_c(b(x) - h(x)) = b(c) - h(c) = z - z = 0.$$

By Theorem 4.44, the irreducible polynomial $g(x)$ must divide $b(x) - h(x)$; thus $b(x) - h(x) = g(x)a(x)$ for some $a(x)$, and so $b(x) = h(x) + g(x)a(x)$, proving our assertion. With the element z in $F[c]$, we now associate this collection of elements of $F[x]$:

$$T(z) = \{b(x) : b(x) \in F[x], \theta_c(b(x)) = z\}. \tag{5.3}$$

In view of the argument just given, we may describe this set as

$$T(z) = \{a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + g(x)a(x) : a(x) \in F[x]\}, \tag{5.4}$$

and where the a_i are the coefficients appearing in 5.1. Note that the polynomial $g(x)$ is characterized as the nonzero polynomial of least degree in the subset $T(0)$. This point will arise again later. We have now produced a collection of subsets of $F[x]$ indexed by the elements of $F[c]$. What is significant here is that these subsets may be defined once the irreducible polynomial $g(x)$ is known; there is no particular need to use the field $F[c]$ to construct the subsets.

There is more to the definition of $F[c]$ than simply the set of its elements; we must consider the operations of addition and multiplication. These operations are reflected by the operations in $F[x]$ as we see in the following computation.

Let $z, w \in F[c]$, and let $u(x)$ be any element in the set $T(z)$, $v(x)$ any element in the set $T(w)$. Then $u(x) + v(x)$ is in the set $T(z + w)$, $u(x)v(x)$ is in the set $T(zw)$. To see this we first note that $u(c) = z$ and $v(c) = w$ and also that $T(z + w)$ is the set of all elements of $F[x]$ which give $z + w$ when evaluated at c. We compute

$$\theta_c(u(x) + v(x)) = u(c) + v(c) = z + w,$$

and so $u(x) + v(x) \in T(z + w)$. A similar argument yields $u(x)v(x) \in T(zw)$. The point of this computation is that the operations defined on $F[c]$ can be defined entirely in terms of the operations in $F[x]$. This gives all the ideas necessary to carry out the construction we have in mind. We shall make the definitions using an arbitrary polynomial rather than an irreducible polynomial. Later we shall restrict our attention to the case of an irreducible polynomial.

The Construction

Let $g(x)$ be a nonzero polynomial in $F[x]$; for each polynomial $h(x) \in F[x]$ define a subset $[h(x)]$ of $F[x]$ by the rule

5.5
$$[h(x)] = \{h(x) + g(x)a(x) : a(x) \in F[x]\}.$$

These subsets are the equivalence classes of an equivalence relation which can be defined on $F[x]$ as follows:

5.6 Definition. Let $g(x)$ be a fixed nonzero polynomial in $F[x]$. For two polynomials $m(x)$ and $n(x)$ in $F[x]$, the phrase "$m(x)$ ***is congruent to*** $n(x)$ ***modulo*** $g(x)$" means $g(x)$ divides $m(x) - n(x)$; we indicate that this holds by writing $m(x) \equiv n(x) \pmod{g(x)}$.

Let us show that this is indeed an equivalence relation according to Definition 1.8. Clearly $g(x)$ divides $m(x) - m(x) = 0$ so that $m(x) \equiv m(x) \pmod{g(x)}$. If $m(x) \equiv n(x) \pmod{g(x)}$ then $g(x)$ divides $m(x) - n(x)$; thus $g(x)$ divides $n(x) - m(x)$ and so $n(x) \equiv m(x) \pmod{g(x)}$. Finally suppose

$$m(x) \equiv n(x)\ (\mathit{mod}\ g(x)), \quad m(x) - n(x) = g(x)a(x);$$
$$n(x) \equiv p(x) \pmod{g(x)}, \quad n(x) - p(x) = g(x)b(x).$$

Then by adding the two equations we obtain

$$m(x) - p(x) = g(x)(a(x) + b(x))$$

and so

$$m(x) \equiv p(x) \pmod{g(x)}.$$

Thus the three conditions in the definition of an equivalence relation are met. As an illustration of this congruence relation let us take $g(x) = x^2 - 2$ in $Q[x]$. Then we have

$$x^3 + 3 \equiv 2x + 3 \pmod{x^2 + 2}$$

because

$$x^3 + 3 - (2x + 3) = x^3 - 2x = x(x^2 - 2) = xg(x).$$

Similarly $x^2 \equiv 2 \pmod{g(x)}$.

Next let us verify that in the general case, the equivalence classes of this relation are those defined in 5.5. If $h(x)$ is any element of $F[x]$, the equivalence class of $h(x)$ is the set of all polynomials $f(x)$ for which $h(x) \equiv f(x) \pmod{g(x)}$; that is, $h(x) - f(x) = g(x)a(x)$ for some $a(x)$. It follows that $f(x) = h(x) - g(x)a(x)$, so $f(x) \in [h(x)]$. It is also obvious that every element in $[h(x)]$ is congruent to $h(x)$ modulo $g(x)$, so we see indeed that $[h(x)]$ is the equivalence class containing $h(x)$.

Next, let us observe that each class $[h(x)]$ contains one and only one polynomial which equals 0 or has degree less than the degree of $g(x)$. By the division algorithm, we have $h(x) = g(x)q(x) + r(x)$ with either $r(x) = 0$ or $\deg r(x) < \deg g(x)$. Since $r(x) \equiv h(x) \pmod{g(x)}$, it follows that $r(x) \in [h(x)]$. Moreover, if $s(x)$ is another polynomial which is either 0 or has degree less than $\deg g(x)$ and if $r(x) \equiv s(x) \pmod{g(x)}$, then $g(x)$ divides $r(x) - s(x)$. Since this is zero or has degree less than $\deg g(x)$, it must be that $s(x) = r(x)$. One way of interpreting this fact is to say that the totality of equivalence classes is a set in one-to-one correspondence with the elements of the form $a_0 + a_1x + \cdots + a_{n-1}x^{n-1}$ if $n = \deg g(x)$.

In our example with $g(x) = x^2 - 2$, the class $[x^3 + 3]$ containing $x^3 + 3$ is the same class as that containing $2x + 3$; moreover $2x + 3$ is the only polynomial in $[x^3 + 3]$ having degree less than $\deg g(x) = 2$. Furthermore, every class contains a polynomial of the form $a + bx$, with $a, b \in \mathbf{Q}$.

Let us use the notation $F[x]/g(x)$ to denote the set of all equivalence classes of elements of $F[x]$ subject to the relation of congruence modulo $g(x)$.

The next step in this construction is to define addition and multiplication of equivalence classes. The rules are based upon the addition and multiplication in the ring $F[x]$. Let $[u(x)]$ and $[v(x)]$ be any two elements of

$F[x]/g(x)$. Addition and multiplication are defined by the rules

5.7
$$[u(x)] + [v(x)] = [u(x) + v(x)],$$
$$[u(x)] \cdot [v(x)] = [u(x)v(x)].$$

A precaution must be taken before we declare that we have defined two operations on $F[x]/g(x)$. It is necessary to show that addition and multiplication are well-defined. This point arose in the discussion of the field of quotients in Section 3.5. We repeat the point for emphasis. The element $u(x)$ is just one of many elements in $[u(x)]$; yet we used that one element in a definition which must apply for the entire set. It is necessary to show that any other choice of element from the set $[u(x)]$ will not affect the sum or product. The crucial information is contained in the following lemma.

5.8 Lemma. *If* $u(x) \equiv u_1(x)$ *(mod* $g(x)$*) and* $v(x) \equiv v_1(x)$ *(mod* $g(x)$*), then*

(*i*) $u(x) + v(x) \equiv u_1(x) + v_1(x) \pmod{g(x)}$,

(*ii*) $u(x)v(x) \equiv u_1(x)v_1(x) \pmod{g(x)}$.

PROOF. To prove these properties, we observe that $u(x) \equiv u_1(x) \pmod{g(x)}$ means that there is a polynomial $h(x)$ with $u(x) = u_1(x) + g(x)h(x)$. Similarly, we have $v(x) = v_1(x) + g(x)k(x)$ for some polynomial $k(x)$. It follows that

$$u(x) + v(x) = u_1(x) + v_1(x) + g(x)(h(x) + k(x));$$

this implies statement (i). To prove the second statement, we observe

$$u(x)v(x) = u_1(x)v_1(x) + g(x)(k(x)u_1(x) + h(x)v_1(x) + g(x)h(x)k(x)),$$

which implies (ii).

The lemma may be restated as follows:

If $[u(x)] = [u_1(x)]$ and $[v(x)] = [v_1(x)]$, then
$[u(x)] + [v(x)] = [u_1(x)] + [v_1(x)]$ and
$[u(x)][v(x)] = [u_1(x)][v_1(x)]$.

This shows that addition and multiplication of elements in $F[x]/g(x)$ are well-defined.

It is now easy to prove the following result.

5.9 Theorem. *The set* $F[x]/g(x)$ *of all equivalence classes of* $F[x]$ *modulo* $g(x)$ *is a commutative ring with unity when addition and multiplication are defined by 5.7.*

PROOF. All of the axioms for a commutative ring follow for $F[x]/g(x)$ from the corresponding axioms for the ring $F[x]$. We give just one illustration, the associative law for multiplication. If $u(x)$, $v(x)$, $w(x) \in F[x]$, then

$$\begin{aligned}([u(x)][v(x)])[w(x)] &= [u(x)v(x)][w(x)]\\ &= [(u(x)v(x))w(x)]\\ &= [u(x)(v(x)w(x))]\\ &= [u(x)][v(x)w(x)]\\ &= [u(x)]([v(x)][w(x)]).\end{aligned}$$

In a similar way the remaining properties may be verified. These steps are left as an exercise for the reader. In particular, notice that the unity of $F[x]/g(x)$ is $[1]$ and the zero element is $[0]$.

As an illustration of the operations, let us consider the ring $\mathbf{Q}[x]/(x^2 - 2)$. We have already pointed out that in this example every class is equal to $[a + bx]$ for some suitable elements $a, b \in \mathbf{Q}$. Here is a sample product:

$$[3 + 5x][-2 + 2x] = [-6 - 4x + 10x^2] = [14 - 4x].$$

This last equality holds because

$$-6 - 4x + 10x^2 = 10(x^2 - 2) + 14 - 4x,$$

and so

$$-6 - 4x + 10x^2 \equiv 14 - 4x \pmod{x^2 - 2}.$$

Of course when doing calculations in the ring $F[x]/g(x)$, we may always replace the class of a polynomial $u(x)$ by the class of its remainder after division by $g(x)$.

Now we come to an important feature of this construction.

5.10 Theorem. *If F is a field, then the ring $F[x]/g(x)$ is a field if and only if $g(x)$ is an irreducible polynomial in $F[x]$.*

PROOF. First suppose that $g(x)$ is not an irreducible polynomial. Then there exist $u(x)$ and $v(x)$ in $F[x]$ with $g(x) = u(x)v(x)$ and neither $u(x)$ nor $v(x)$ has degree zero. In particular, $g(x)$ does not divide either $u(x)$ or $v(x)$. This means that $u(x) \not\equiv 0 \pmod{g(x)}$ and $v(x) \not\equiv 0 \pmod{g(x)}$; this translates to $[u(x)] \neq [0]$ and $[v(x)] \neq [0]$. But

$$[u(x)][v(x)] = [u(x)v(x)] = [g(x)] = [0].$$

Hence we have found the product of these two nonzero elements equals zero. This means that $F[x]/g(x)$ is not an integral domain and so surely is not a field.

Now let us suppose that $g(x)$ is an irreducible polynomial. We must show that every nonzero element in $F[x]/g(x)$ has a multiplicative inverse. Let $u(x)$ be a polynomial with $[u(x)] \neq [0]$. Then $u(x) \not\equiv 0 \pmod{g(x)}$ and so $g(x)$ does not divide $u(x)$. Our assumption that $g(x)$ is irreducible implies that $g(x)$ is a prime polynomial and so it follows that the g.c.d. of $u(x)$ and $g(x)$ is 1. By Theorem 4.26, there exist $a(x)$ and $b(x)$ in $F[x]$ such that $u(x)a(x) + g(x)b(x) = 1$. Now use the equivalence classes to interpret this as

$$\begin{aligned}[1] &= [u(x)a(x) + g(x)b(x)] = [u(x)a(x)] + [g(x)b(x)]\\ &= [u(x)][a(x)] + [0]\\ &= [u(x)][a(x)].\end{aligned}$$

Thus the class $[u(x)]$ has an inverse, namely, the class $[a(x)]$. This proves that when $g(x)$ is irreducible, every nonzero class in $F[x]/g(x)$ has an inverse; that is, $F[x]/g(x)$ is a field.

The construction just given starts with a field F and an irreducible polynomial $g(x) \in F[x]$, and produces a field $F[x]/g(x)$. Let us now show how this field may be interpreted as an extension field of F containing a root of $g(x)$.

Define a mapping θ from $F[x]$ to $F[x]/g(x)$ by the rule

$$\theta(f(x)) = [f(x)].$$

The definition of addition and multiplication given in 5.7 implies that θ is a homomorphism of $F[x]$ into $F[x]/g(x)$. Clearly this map is in fact onto $F[x]/g(x)$. The function θ is used as a device to identify certain elements of $F[x]/g(x)$ with elements of F. Let F' be the set of classes in $F[x]/g(x)$ which contain an element of F. For each $b \in F$, $\theta(b) = [b]$ is in F'. Thus θ establishes a mapping from F to F'. It is easily verified that this mapping is one to one and onto; that is, θ establishes an isomorphism between F and F'. Thus $F[x]/g(x)$ contains a subfield isomorphic with F and we shall use this isomorphism to identify the elements of F' and F. In place of $[b]$ in F', we shall simply write b; in place of $[bx] = [b][x]$, we write $b[x]$. More generally,

$$[b_0 + b_1x + \cdots + b_kx^k] = b_0 + b_1[x] + \cdots + b_k[x^k].$$

Now we go one step further and introduce a symbol for $[x]$. Let $c = [x]$. Then we have $c^k = [x]^k = [x^k]$ and so a typical element of $F[x]/g(x)$ has the form

$$[b_0 + b_1x + \cdots + b_kx^k] = b_0 + b_1c + \cdots + b_kc^k.$$

With this agreement on notation, the map θ is nothing more than the evaluation map at c. Let us see why this is the case. Set $L = F[x]/g(x)$ so L is a field. The map θ has values in L and for

$$h(x) = b_0 + b_1x + \cdots + b_kx^k,$$

we have

$$\begin{aligned}\theta(h(x)) &= [h(x)] = b_0 + b_1[x] + \cdots + b_k[x^k] \\ &= b_0 + b_1 c + \cdots + b_k c^k \\ &= h(c).\end{aligned}$$

Thus θ is the evaluation at c. Observe also that the class of $g(x)$ is the zero class, $[g(x)] = [0]$, so

$$\theta(g(x)) = 0 = g(c).$$

Thus c is a root of $g(x)$ in the field L. We generalize this result in the following.

5.11 Theorem. *Let F be any field and $f(x)$ a polynomial of positive degree in $F[x]$. Then there is a field L containing F and containing a root of $f(x)$.*

PROOF. Let $f(x) = g(x)h(x)$ with $g(x)$ an irreducible factor of positive degree. Then the field $L = F[x]/g(x)$ contains F and a root c of $g(x)$. Then $f(c) = g(c)h(c) = 0$, so c is also a root of $f(x)$.

We can extend this result slightly to obtain the following.

5.12 Theorem. *Let F be any field and $f(x)$ a polynomial of positive degree n in $F[x]$. Then there is a field extension E of F such that*

5.13 $$f(x) = a(x - c_1)(x - c_2)\cdots(x - c_n)$$

with a = leading coefficient of $f(x)$ and $c_1, c_2, \ldots, c_n$ elements of E.

PROOF. By Theorem 5.11, there is a field extension L_1 of F which contains a root c_1 of $f(x)$. By the Factor Theorem, $f(x) = (x - c_1)h(x)$, with $h(x)$ in $L_1[x]$. If $h(x)$ has positive degree, repeat this using $h(x)$ in place of $f(x)$ and L_1 in place of F. There is a field L_2 containing L_1 and a root c_2 of $h(x)$ so that

$$f(x) = (x - c_1)(x - c_2)h_2(x)$$

with $h_2(x)$ in $L_2[x]$. After this is repeated sufficiently, the factorization 5.13 is obtained with the c_i in some field containing F.

If $f(x)$ has the factorization 5.13 in a field L containing F, then L is called a *splitting field* for the polynomial $f(x)$ over F. For example, $\mathbf{R}[i] = \mathbf{C}$ is a splitting field for $f(x) = x^2 + 1$. For another example, consider the polynomial $f(x) = x^3 - 1$ in $\mathbf{Q}[x]$. The complex number $w = (-1 + i\sqrt{3})/2$ is a root of $f(x)$ (check this!) as is also w^2 and $w^3 = 1$. Thus

$$x^3 - 1 = (x - 1)(x - w)(x - w^2)$$

and $\mathbf{Q}[w]$ is a splitting field for $f(x)$ over $\mathbf{Q}$.

If $f(x)$ has the factorization 5.13, then the smallest field containing F and each of the roots, c_i, is denoted by $F[c_1,\ldots,c_n]$. It can be shown that this field is uniquely determined up to isomorphism by F and $f(x)$. We will not carry out the proof of this result. It may be considered as the beginning of the subject known as Galois theory. Some further illustrations of Galois theory will be given in the section on finite fields.

EXERCISES

1. Find the unique element in each equivalence class z of $\mathbf{Q}[x] \bmod g(x)$ which has degree less than the degree of $g(x)$:
   (i) $z = [x^3 - 3x + 2]$, $g(x) = x^2 - 2$;
   (ii) $z = [x^4 + x^2 + 1]$, $g(x) = x^2 + x + 1$;
   (iii) $z = [x^2 + 4][x^3 - 3x]$, $g(x) = x^2 + 1$;
   (iv) $z = [x^2 - 3x + 2][x - 2] + [2x^3 - x + 1]$, $g(x) = x^2 - 3$.
2. In the field $\mathbf{Q}[x]/g(x)$, find the inverse of z for each z and $g(x)$:
   (i) $z = [x + 1]$, $g(x) = x^2 - 2$;
   (ii) $z = [x + 1]$, $g(x) = x^2 + 2$;
   (iii) $z = [3x^2 - 5]$, $g(x) = x^3 - 5$.
3. Let $\mathbf{R}$ denote the field of real numbers. Give an explicit isomorphism of $\mathbf{R}[x]/(x^2 + 1)$ with the field $\mathbf{C}$ of complex numbers.
4. Let $g(x)$ be any irreducible polynomial of degree 2 in $\mathbf{R}[x]$. Write $g(x) = x^2 + bx + c$; then give an explicit isomorphism of $\mathbf{R}[x]/g(x)$ with $\mathbf{C}$.
5. Let E be a field containing the rational numbers $\mathbf{Q}$ and an element $c \neq 1$ which is a root of the polynomial $f(x) = x^5 - 1$. Show that the subfield $\mathbf{Q}[c]$ of E contains 5 roots of $f(x)$ and hence that $\mathbf{Q}[c]$ is a splitting field of $f(x)$.
6. Let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ and suppose $f(x)$ has the factorization 5.13 with the c_i in some splitting field. Prove
   (i) $c_1 + c_2 + \cdots + c_n = -a_{n-1}$;
   (ii) $c_1c_2\cdots c_n = (-1)^n a_0$.
7. Let $g(x)$ be an irreducible polynomial of degree 2 in $F[x]$ and let L be a field which contains a root c of $g(x)$. Show $F[c]$ is a splitting field for $g(x)$.
8. The polynomial $g(x) = x^2$ is not irreducible in $\mathbf{Q}[x]$, so $\mathbf{Q}[x]/x^2$ is not a field. Find all the zero divisors in this ring. Find all the elements which have a multiplicative inverse. Prove that every nonzero element is either a zero divisor or has an inverse.

5.2 IDEALS

In this section, the construction of the ring $F[x]/g(x)$ as a homomorphic image of $F[x]$ is put into a more general context. We shall describe how to obtain homomorphic images of any ring. Even though we are primarily interested in commutative rings, the definitions are given to allow for noncommutative rings as well.

Let us begin with the following definition.

5.14 Definition. Let A be a subring of the ring R. Then

(i) A is said to be a ***right ideal*** in R if A is closed with respect to multiplication on the right by elements of R (if $a \in A$ and $r \in R$, then $ar \in A$).
(ii) A is said to be a ***left ideal*** in R if A is closed with respect to multiplication on the left by elements of R (if $a \in A$ and $r \in R$, then $ra \in A$).
(iii) A is said to be an ***ideal*** in R if it is both a right ideal in R and a left ideal in R (that is, it is closed with respect to multiplication on either side by elements of R).

We shall primarily be interested in ideals in a ring although the study of right (or left) ideals plays an important role in more advanced ring theory.

In any ring R, the subring consisting only of the zero element is clearly an ideal, and the entire ring R is also an ideal. These two ideals are often called *trivial ideals*. Another simple observation is that if R has a unity e and a right ideal A in R contains an element a with a multiplicative inverse, then $A = R$. For if $a \in A$ and $aa^{-1} = e$, then $e \in A$ and $ex = x \in A$ for every x in R. Clearly, a similar result holds for left ideals and ideals.

As a first example of an ideal (other than the trivial ones mentioned above), we may observe that the subring E of even integers in the ring $\mathbf{Z}$ is an ideal in $\mathbf{Z}$. This is true since the product of an even integer by an arbitrary integer is an even integer.

The subring $\mathbf{Z}$ of integers in the ring $\mathbf{Q}$ of rational numbers is not an ideal in $\mathbf{Q}$ since, for example, $3 \in \mathbf{Z}$ and $\frac{1}{2} \in \mathbf{Q}$, but $3 \cdot \frac{1}{2} \notin \mathbf{Z}$.

In order to give an example of a right ideal (or left ideal) which is not an ideal, we must clearly have a noncommutative ring. It is easy to verify that the subring $\{a, b\}$ of the ring K of Example 6 of Section 2.3 is a left ideal, but not a right ideal, in K. A more significant example is the following. Let $M_2(\mathbf{Z})$ be the ring of all two-by-two matrices over the integers (Example 9 of Section 2.3). Then the set of all elements of $M_2(\mathbf{Z})$ of the form

$$\begin{bmatrix} x & y \\ 0 & 0 \end{bmatrix}, \qquad x, y \in \mathbf{Z},$$

is a right ideal but not a left ideal; the set of all elements of $M_2(\mathbf{Z})$ of the form

$$\begin{bmatrix} x & 0 \\ y & 0 \end{bmatrix}, \qquad x, y \in \mathbf{Z},$$

is a left ideal but not a right ideal; and the set of all elements of $M_2(\mathbf{Z})$ of the form

$$\begin{bmatrix} x & y \\ z & t \end{bmatrix},$$

where x, y, z, and t are *even* integers, is an ideal in $M_2(\mathbf{Z})$. We leave to the reader the verification of these statements.

Throughout the rest of this section we shall consider *commutative* rings only, and hence there will be no distinction between ideals and right (or left) ideals. Moreover, for our purposes the most important case is that in which the ring has a unity. Accordingly, let S be a commutative ring with unity e. If $a \in S$, let

$$A = \{as \mid s \in S\},$$

and let us verify that A is an ideal in S. If $s, t \in S$, then $as + at = a(s + t)$ and $(as)t = a(st)$, so that A is clearly closed with respect to addition and with respect to multiplication by arbitrary elements of S. Moreover, $-(as) = a(-s) \in A$, and hence additive inverses of elements of A are elements of A. Thus, A is indeed an ideal in S. This conclusion holds whether or not S has a unity, but the presence of a unity now assures us that $a \in A$ since $a = ae$. It is customary to denote this ideal A by (a). As an example of this notation, the ideal of all even integers in the ring $\mathbf{Z}$ would be denoted by (2). Similarly, the ideal consisting of all multiples of 3 would be denoted by (3), and so on.

Some convenient terminology is introduced in the following definition.

5.15 Definition. Let S be a commutative ring with unity. If $a \in S$, an ideal of the form

$$(a) = \{as \mid s \in S\}$$

is called a ***principal ideal.*** It is also called the ***principal ideal generated by a*** and a is called a ***generator*** of the ideal.

The next theorem shows that in certain important rings there are no ideals except principal ideals.

5.16 Theorem. *If R is a Euclidean domain, then every ideal of R is principal.*

PROOF. Let d be the Euclidean function for R and let A be any ideal of R. If A consists of only the zero element, then A is principal with $A = (0)$.

Suppose then that A is not the zero ideal. Then the set

$$M = \{d(a) \mid a \in A, a \neq 0\}$$

is not the empty set. Since M consists of nonnegative integers, the well-ordering property implies there is a least element in M. Select $b \in A$ so that $d(b)$ is the least element of M; then $d(b) \leq d(a)$ for each $a \in A$ with $a \neq 0$. We now prove that A is the principal ideal generated by b. To do this, we must show that every element of A is divisible by b. Take any $a \in A$; by the division algorithm we have $a = bq + r$ with either $r = 0$ or $d(r) < d(b)$. This second alternative is not possible for the following reason. The element $r = a - bq$ is an element of A, since both a and b are in A. If r is not zero, then $d(r)$ is in M and is smaller than the smallest element of M, namely, $d(b)$. This is an impossible situation, so in fact $r = 0$ and $a = bq$; this shows every element of A is a multiple of b and so $A = (b)$.

We have proved earlier that the rings $\mathbf{Z}$ and $F[x]$ are Euclidean rings. The theorem gives us the following.

5.17 Corollary. *Every ideal is a principal ideal in*

(*i*) *The ring* $\mathbf{Z}$ *of integers,*
(*ii*) *The polynomial ring* $F[x]$ *over a field* F.

Let us observe that the rings $\mathbf{Z}$ and $F[x]$ are quite special rings, and the result just proved for these rings does not hold in general. We shall now emphasize this fact by giving an example of an ideal which is not a principal ideal. For our ring we take $\mathbf{Z}[x]$, the ring of polynomials over $\mathbf{Z}$. It may be verified that the set T of all elements of $\mathbf{Z}[x]$ with *even* constant terms is an ideal in $\mathbf{Z}[x]$. However, we shall show that T is not a principal ideal. Suppose, on the contrary, that $T = (f(x))$. Since $2 \in T$, we must have $2 = f(x)g(x)$ for some element $g(x)$ of $\mathbf{Z}[x]$. It follows that both $f(x)$ and $g(x)$ are of degree zero, that is, they are integers. Now $f(x) \neq \pm 1$ since this would imply that $(f(x)) = \mathbf{Z}[x]$, which is false. Accordingly, we must have $f(x) = \pm 2$, $g(x) = \pm 1$. However, there are many elements of T (for example, $x + 2$) which are not multiples of ± 2. We have obtained a contradiction, and T is therefore not a principal ideal in $\mathbf{Z}[x]$.

EXERCISES

1. If A and B are ideals (right ideals, left ideals) in a ring R, prove that $A \cap B$ is an ideal (right ideal, left ideal) in R. Generalize to any number of ideals.
2. If A and B are ideals (right ideals, left ideals) in a ring R, let us define

$$A + B = \{a + b \mid a \in A, b \in B\}.$$

Prove that $A + B$ is an ideal (right ideal, left ideal) in R and that $A \subseteq A + B$ and $B \subseteq A + B$.

3. Let $A = (x^2 - x - 6)$ and $B = (x^2 - 4)$ be principal ideals in $\mathbf{Q}[x]$. Then $A + B$ and $A \cap B$ are principal ideals. Find generators for each of them.
4. Let s and t be nonzero integers having g.c.d. equal to d. Prove that in the ring $\mathbf{Z}$, $(s) + (t) = (d)$.
5. Verify that the only ideals in a field F are the two trivial ideals (0) and F.
6. Let R be a commutative ring with unity such that the only ideals of R are the two trivial ideals (0) and R. Prove R is a field.

5.3 IDEALS, CONGRUENCE RELATIONS, AND HOMOMORPHISMS

We begin by using an ideal A of a ring R to define an equivalence relation.

5.18 Definition. If A is an ideal of a ring R and if $u, v \in R$, we define the relation "***u is congruent to v modulo A***" to mean $u - v \in A$. We write $u \equiv v \pmod{A}$ to indicate that this holds.

The relation of congruence modulo A is an equivalence relation; the three conditions of Definition 1.8 are easily checked and will be left as an exercise. It should be observed that in the case $R = F[x]$ and $A = (g(x))$, the relation just defined, $u \equiv v \pmod{A}$, is exactly the same as the relation $u \equiv v \pmod{g(x)}$ defined in 5.6.

The use of the symbol "$\equiv$" is intended to suggest "equality" since many properties which are taken for granted when using "$=$" are valid also for "$\equiv$." We make this explicit with examples.

5.19 Theorem. *Let u, v, u_1, v_1 be elements of the ring R and let A be an ideal of R. If*

$$u \equiv u_1 \pmod{A} \quad \text{and} \quad v \equiv v_1 \pmod{A},$$

then

(i) $u + v \equiv u_1 + v_1 \pmod{A}$,
(ii) $uv \equiv u_1 v_1 \pmod{A}$.

PROOF. The condition $u \equiv u_1 \pmod{A}$ implies there is an element $a \in A$ with $u = u_1 + a$; similarly $v = v_1 + b$ with $b \in A$. Then we obtain the equations

$$u + v = u_1 + v_1 + (a + b),$$
$$uv = u_1v_1 + (av_1 + u_1b + ab).$$

The definition of an ideal implies that $(a + b)$ and $(av_1 + u_1b + ab)$ are both elements of A. Thus statements (i) and (ii) are valid.

We continue the analogy with the construction of the ring $F[x]/g(x)$. Let R/A denote the set of equivalence classes of R modulo the ideal A. This symbol may be read as "R modulo A" or more simply "R mod A." Our plan is to define operations on R/A to make it a ring. Our previous notation would dictate the use of the symbol $[u]$ for the equivalence class of the element u. We shall, however, introduce a change of notation which will be more suggestive. The set $[u]$ consists of all elements v in R which satisfy $u \equiv v \pmod{A}$; this is the same as saying $v - u$ is in A. Putting this yet another way, we see that the equivalence class $[u]$ consists of all v such that $v = u + (v - u) = u + a$ for an element a in A. This suggests the notation

5.20 $$u + A = \{u + a : a \in A\}.$$

The set $u + A$ is called a ***coset of*** A or, more precisely, the coset of A containing u. In order to gain some familiarity with this notation, we repeat some properties of cosets which are consequences of the properties 1.11 of equivalence classes:

5.21
(i) $u + A = v + A$ if and only if $u - v \in A$;
(ii) $(u + A) \cap (v + A) \neq \varnothing$ if and only if $u + A = v + A$.

In particular, $0 + A = A$ is a coset of A and $u + A = A$ if and only if $u \in A$.

5.22 Theorem. *Let A be an ideal of the ring R. Operations of addition and multiplication are defined on the set R/A of cosets of A by the rules*

5.23 $$(u + A) + (v + A) = (u + v) + A,$$

5.24 $$(u + A)(v + A) = uv + A.$$

With these operations, R/A is a ring. Moreover the mapping $\theta : R \to R/A$ defined by $\theta(u) = u + A$ is a homomorphism of R onto R/A.

PROOF. As we have seen several times previously, the main point is to show that addition and multiplication are well-defined. We carry this out only for addition. Suppose $u + A = u_1 + A$ and that $v + A = v_1 + A$. Then $u \equiv u_1 \pmod{A}$ and $v \equiv v_1 \pmod{A}$ and so by 5.19(i), we conclude $(u + v) \equiv (u_1 + v_1) \pmod{A}$. Thus $(u + v) + A = (u_1 + v_1) + A$;

as a result of this, the sum of the cosets given in 5.23 is the same no matter what elements are used to represent the particular cosets. Similarly, one can use 5.19(ii) to show that multiplication is well-defined by 5.24. At this point we have the set R/A with two operations defined on it. In order to show that R/A is a ring, we use the map $\theta: R \to R/A$ which is defined by $\theta(u) = u + A$. Clearly this maps R onto R/A. The Equations 5.23 and 5.24 can be rewritten as

$$\theta(u + v) = \theta(u) + \theta(v),$$
$$\theta(uv) = \theta(u)\theta(v).$$

The validity of the ring axioms in R now implies the validity of the same axioms in R/A. For example to show the associative law for multiplication in R/A we have

$$\begin{aligned}(u + A)[(v + A)(w + A)] &= \theta(u)[\theta(v)\theta(w)] = \theta(u)\theta(vw)\\ &= \theta(u(vw)) = \theta((uv)w)\\ &= \theta(uv)\theta(w) = [\theta(u)\theta(v)]\theta(w)\\ &= [(u + A)(v + A)](w + A).\end{aligned}$$

In a similar way all the ring axioms are seen to hold for R/A. Thus R/A is a ring and, moreover, this reasoning shows that θ is a homomorphism of R onto R/A.

The ring R/A is called a ***factor ring*** of R.

It is worth noting that $0 + A = A$ is the zero element of R/A. If R has a unity element e, then $e + A$ is the unity of R/A. If the ring R is commutative, then every factor ring R/A is also commutative. The converse is not true, however. The ring R/A may be commutative even though R is not. (See Exercise 6 below.)

Next we shall indicate how this construction of factor rings of R is, in fact, an essential part of the construction of the homomorphic images of R. We begin with a definition.

5.25 Definition. If $\phi: R \to S$ is a homomorphism from the ring R to the ring S, then the set $K = \{r : r \in R \text{ and } \phi(r) = 0\}$ is called the ***kernel*** of ϕ. The kernel of ϕ is denoted by $\ker \phi$.

The relation of this idea with that of the preceding material is given in the following result.

5.26 Theorem. *If $\phi: R \to S$ is a homomorphism of the ring R into the ring S, then $\ker \phi$ is an ideal of R. Moreover, ϕ is a one-to-one mapping if* *and only if* $\ker \phi = (0)$. *If* ϕ *maps* R *onto* S, *then* ϕ *is an isomorphism if and only if* $\ker \phi = (0)$.

PROOF. Let $A = \ker \phi$. In order to show A is an ideal, it is necessary to show two facts. For the first, we assume a and b are elements of A. Then

$$\phi(a + b) = \phi(a) + \phi(b) = 0 + 0 = 0;$$

thus $a + b$ is in the kernel, A. For the second fact assume $a \in A$ and $r \in R$. Then

$$\phi(ar) = \phi(a)\phi(r) = 0 \cdot \phi(r) = 0,$$
$$\phi(ra) = \phi(r)\phi(a) = \phi(r) \cdot 0 = 0.$$

Thus both ar and ra are in A and so A is an ideal of R.

Now suppose ϕ is a one-to-one mapping. If $a \in \ker \phi$, then $\phi(a) = 0 = \phi(0)$. The one-to-one property implies that $a = 0$; hence $\ker \phi = (0)$. To go the other way, suppose that $\ker \phi = (0)$ and that a and b are elements of R such that $\phi(a) = \phi(b)$. Then $\phi(a - b) = 0$ and so $a - b$ is in $\ker \phi = (0)$; this implies $a - b = 0$. It follows that the mapping is one-to-one. Moreover, if it is known in advance that ϕ maps R onto S, then ϕ is an isomorphism if and only if it is one-to-one, that is, if and only if $\ker \phi = (0)$. Hence all parts of the theorem have been proved.

The final result in this section gives the fundamental relationship between factor rings and homomorphic images of R.

5.27 Theorem (THE FIRST ISOMORPHISM THEOREM). *Let* ϕ *be a homomorphism of the ring* R *onto the ring* S *with kernel* A. *Then* S *is isomorphic with the factor ring* R/A. *More precisely, the map* $\alpha: R/A \to S$ *defined by*

$$\alpha(u + A) = \phi(u) \tag{5.28}$$

is a well-defined isomorphism from R/A *onto* S.

PROOF. We have shown in 5.26 that A is an ideal and so the factor ring R/A exists. In order to show that the map α given by 5.28 is well-defined, it is necessary to show that whenever $u + A = v + A$, then $\phi(u) = \phi(v)$. This follows because $v = u + a$ for some $a \in A$, and

$$\phi(v) = \phi(u + a) = \phi(u) + \phi(a) = \phi(u)$$

since a is in the kernel. Thus α is a function from R/A to S. It is necessary to verify that it is, first, a homomorphism, second, a one-to-one map, and third, an onto map. To see that α is a homomorphism we compute

$$\begin{aligned}\alpha((u + A) + (v + A)) &= \alpha(u + v + A) = \phi(u + v)\\ &= \phi(u) + \phi(v)\\ &= \alpha(u + A) + \alpha(v + A).\end{aligned}$$

This shows that α preserves addition; a similar computation shows that α preserves multiplication, and so α is a homomorphism. Next we show that α is one-to-one; to do this it is sufficient by Theorem 5.26 to show that $\ker \alpha = (0)$. Let $u + A$ be an element of $\ker \alpha$. Then

$$0 = \alpha(u + A) = \phi(u)$$

which means that u is in $\ker \phi = A$; it follows that $u + A = 0 + A = A$, which is the zero of the ring R/A. Hence $\ker \alpha = (0)$. Finally we show α is onto. Take any element $s \in S$. We assumed that ϕ was onto, so there is some $r \in R$ with $\phi(r) = s$. Then $\alpha(r + A) = \phi(r) = s$ and this proves α is onto. All parts of the theorem hold.

This theorem is called the First Isomorphism Theorem for Rings because it plays a fundamental role in the study of homomorphism. It will be referred to frequently. The theorem may be interpreted as stating that if isomorphic rings are not considered as different, then the only homomorphic images of a ring R are the factor rings R/A, where A is an ideal of R.

Let us give a few illustrations of the use of this theorem. Suppose F is a field and $\phi: F \to S$ is a homomorphism of F onto S. In order to determine some information about S, we need to know something about the ideals of F. Every nonzero element of F has an inverse, so any nonzero ideal of F contains the unity element and hence contains every element of F. Thus F has only the two trivial ideals. It follows that there are only two possibilities for $A = \ker \phi$. If $A = (0)$, then by Theorem 5.26, ϕ is an isomorphism; that is, S is a field isomorphic with F. The other alternative is that $A = F$ in which case $F/A = F/F$ is a ring with only one element since every coset $f + F = 0 + F$. Thus 5.27 implies S is a ring with only one element.

Suppose $\phi: F[x] \to S$ is a homomorphism and both F and S are fields. Then 5.27 implies that S is isomorphic with $F[x]/(g(x))$ for some polynomial $g(x)$. Then $g(x)$ cannot be the zero polynomial because $F[x]/(0)$ is isomorphic to $F[x]$, and this is not a field. If $g(x)$ has degree zero, then $g = g(x)$ is a nonzero element of F and the ideal $(g(x))$ is all of $F[x]$ because it contains $g \cdot g^{-1} = 1$. Thus $S = 0$. Finally if $g(x)$ has degree at least one, then $g(x)$ must be an irreducible polynomial in $F[x]$. Thus S is isomorphic to a field as constructed earlier in Section 5.1. S must contain an isomorphic copy of F since $\phi(F)$ is isomorphic to F and is in S. Thus this argument shows that the previous construction has produced all the fields which can be obtained as a homomorphic image of $F[x]$. In the next section, we shall apply Theorem 5.27 to discuss all the homomorphic images of the ring of integers.

EXERCISES

1. Let $A = (g(x))$ and $B = (h(x))$ be two ideals of $\mathbf{Q}[x]$. Show that $A \subseteq B$ if and only if $h(x) \mid g(x)$.
2. Let $g(x) = (x + 1)^2(x - 2)^3$. What are all the ideals of $\mathbf{Q}[x]$ which contain $g(x)$? (There are 12 of them.)
3. Let $\phi: R \to S$ be a homomorphism of the ring R onto the ring S. Define a map $\theta: R[x] \to S[x]$ by
$$\theta(r_0 + r_1x + \cdots + r_nx^n) = \phi(r_0) + \phi(r_1)x + \cdots + \phi(r_n)x^n.$$
Show that θ is a homomorphism of $R[x]$ onto $S[x]$. Describe the kernel of θ in terms of the kernel of ϕ.
4. (a) Let A and B be ideals of a ring R with $A \subseteq B$. Show that B/A, the set of cosets of the form $b + A$ with $b \in B$, is an ideal of the ring R/A.
   (b) Since B/A is an ideal of R/A, the factor ring $(R/A)/(B/A)$ is defined. Show that $(R/A)/(B/A)$ is isomorphic to R/B. Carry this out as follows: define a map $\alpha: R/A \to R/B$ by the rule $\alpha(u + A) = u + B$. Show that α is a homomorphism of R/A onto R/B with kernel B/A. Apply Theorem 5.27. (This result is called the *Second Isomorphism Theorem for Rings*.)
5. Let A be an ideal of the ring R. Prove that R/A is a commutative ring if and only if $rs - sr \in A$ for every $r, s \in R$.
6. Let R be the ring of all two by two matrices of the form
$$r(x, y, z) = \begin{bmatrix} x & 0 \\ y & z \end{bmatrix}, \qquad x, y, z \in \mathbf{Z}.$$
Let A be the subset of R consisting of all elements $r(0, y, 0)$, $y \in \mathbf{Z}$. Prove (i) A is an ideal of R; (ii) R is noncommutative but R/A is commutative.

5.4 THE INTEGERS MODULO *n*

In this section we examine the homomorphic images of the ring of integers, $\mathbf{Z}$. The proper homomorphic images, that is, those not isomorphic to $\mathbf{Z}$, are all finite rings. These rings are of interest in their own right, but will also be used to prove some facts about divisibility of integers.

It was proved in 5.17 that every ideal in $\mathbf{Z}$ is principal. We consider a nonzero ideal (n). It may be assumed that $n > 0$ since $(n) = (-n)$; that is, every integer which is divisible by n is also divisible by $-n$. The general definition of congruence modulo an ideal given in Definition 5.18 is restated here for the case of integers:

If $a, b \in \mathbf{Z}$, then $a \equiv b \pmod{n}$ if and only if $n \mid (a - b)$.

Congruence mod n is an equivalence relation on $\mathbf{Z}$ whose equivalence classes are defined by

$$[u] = \{v : v \in \mathbf{Z} \text{ and } u \equiv v \pmod{n}\}.$$

Occasionally it will be useful to use the notion of cosets for these classes:

$$u + (n) = \{u + nt : t \in \mathbf{Z}\}.$$

As we have seen in the more general case discussed in the previous section, the sets $[u]$ and $u + (n)$ contain exactly the same integers. The notation $[u]$ has the advantage of being simple to write; the notation $u + (n)$ has the advantage of displaying both the integer u and the integer n. The set of all equivalence classes of integers modulo n has been denoted by $\mathbf{Z}/(n)$ in the previous section; we shall use the simpler notation $\mathbf{Z}_n$ to denote this set. Then $\mathbf{Z}_n$ is a ring with operations

$$[u] + [v] = [u + v], \quad [u] \cdot [v] = [uv].$$

These operations are well-defined and $\mathbf{Z}_n$ is a commutative ring with unity.

Since we have several notations for the same objects, let us consider a few numerical examples using the different notations. Take $n = 5$. Then we have

$$18 \equiv 3 \pmod{5} \quad \text{and} \quad -9 \equiv 1 \pmod{5}$$

because 5 divides $(18 - 3)$ and $(-9 - 1)$. These two congruences are equivalent to the statements

$$[18] = [3] \quad \text{and} \quad [-9] = [1]$$

or

$$18 + (5) = 3 + (5) \quad \text{and} \quad -9 + (5) = 1 + (5).$$

Here are some examples to illustrate the ring operations:

$$\begin{aligned}
&[18] + [7] = [25] = [0], \\
&[13] + [6] = [19] = [4], \\
&[3][2] = [6] = [1], \\
&[2]^{100} = ([2]^4)^{25} = [2^4]^{25} = [16]^{25} = [1]^{25} = [1].
\end{aligned}$$

Every element in $\mathbf{Z}_n$ is a set with many integers. It is useful to select the smallest nonnegative integer in each set in order to deal with smaller numbers. This will also help us determine how many elements there are in $\mathbf{Z}_n$.

For any integer u, apply the division algorithm to obtain

$$u = nq + r, \quad 0 \le r < n.$$

It follows that $[u] = [r]$ since $u - r$ is divisible by n. Thus every class $[u]$ is equal to a class $[r]$ in which $0 \le r < n$. Moreover, any given class contains only one such representative. For suppose that $0 \le s < n$ and $[s] = [r]$. Then n divides $r - s$; the inequalities satisfied by r and s imply that this difference satisfies $-n < r - s < n$. The only integer on this interval which is a multiple of n is 0; thus $r - s = 0$. This means that the complete list of elements in the ring $\mathbf{Z}_n$ is $[0], [1], [2], \ldots, [n-1]$. Thus $\mathbf{Z}_n$ is a ring with exactly n elements. We collect these and a few additional facts in a theorem.

5.29 Theorem. *If n is a positive integer, then the ring $\mathbf{Z}_n$ is a homomorphic image of $\mathbf{Z}$ having exactly n elements* $[0], [1], \ldots, [n-1]$.

PROOF. We have already proved these statements in the remarks above and in the more general case of the previous section. We emphasize that the mapping $\mathbf{Z} \to \mathbf{Z}_n$ defined by $m \to [m]$ (or $m \to m + (n)$) is the homomorphism of $\mathbf{Z}$ onto $\mathbf{Z}_n$. The unity of $\mathbf{Z}_n$ is $[1] = 1 + (n)$.

As an illustration, let us again consider the special case in which $n = 5$. As pointed out above, the five elements of this ring are $[0], [1], [2], [3]$, and $[4]$. The reader may verify the following addition and multiplication tables for this ring. For convenience, we have omitted the brackets and written "k" in place of "$[k]$." This is often done when the context makes the meaning clear.

(+)	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

(·)	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

The Ring $\mathbf{Z}_5$

By examining the multiplication table for this ring we see that this ring has no nonzero divisors of zero, and hence that the ring is an integral domain. The following theorem tells us for just what integers n the ring $\mathbf{Z}_n$ is an integral domain.

5.30 Theorem. *The ring $\mathbf{Z}_n$ is an integral domain if and only if n is a prime integer. Moreover when n is a prime, $\mathbf{Z}_n$ is a field.*

PROOF. Suppose first that n is not a prime. Then there exist integers n_1, n_2 with $n = n_1 n_2$ and $1 < n_1, n_2 < n$. Now in $\mathbf{Z}_n$ we have

$$[n_1]\cdot[n_2] = [n_1 n_2] = [n] = [0].$$

However $[n_1] \neq [0]$, $[n_2] \neq 0$ because n does not divide either n_i. So we have found two nonzero elements of $\mathbf{Z}_n$ with product equal to zero. This shows that $\mathbf{Z}_n$ is not an integral domain.

Now suppose that n is a prime integer. We shall prove that $\mathbf{Z}_n$ is a field; so in particular it is an integral domain. To show $\mathbf{Z}_n$ is a field, it is necessary to show that every nonzero element has a multiplicative inverse. Suppose $[a] \neq [0]$. Then n does not divide the integer a and, since n is a prime, the g.c.d. of n and a is 1. By Theorem 4.32, there exist integers x and y such that $ax + ny = 1$; thus n divides $ax - 1$ and so

$$[1] = [ax] = [a][x]$$

which proves that $[a]$ has an inverse, as required.

This construction produces a new class of fields that is entirely different from the rational numbers, real numbers, or complex numbers which have previously been the fields used in all examples. With these finite fields in mind, we shall apply the results of Chapter 4 in the next section.

The main step in proving $\mathbf{Z}_n$ is a field when n is prime can be applied to identify the elements of $\mathbf{Z}_n$ which have an inverse even when n is not a prime. Recall that in Section 4.4 we called an element of a ring having a multiplicative inverse a *unit*. The notation $U(\mathbf{Z}_n)$ denotes the set of all units of the ring $\mathbf{Z}_n$.

5.31 Theorem. *The element* $[u]$ *of* $\mathbf{Z}_n$ *is a unit if and only if the g.c.d. of u and n equals* 1.

PROOF. The class $[v]$ is the inverse of $[u]$ if and only if there is an integer t such that $uv - 1 = nt$. If we treat u and n as given and $v = x$ and $t = -y$ as unknowns, then the equation looks like $ux + ny = 1$ which has been treated in Section 4.5; it implies this equation has a solution if and only if the g.c.d. of u and n equals 1.

Euler's Phi Function

For each positive integer n, we denote by $\phi(n)$ the number of elements in the set $U(\mathbf{Z}_n)$, the units of $\mathbf{Z}_n$. In view of the last theorem, $\phi(n)$ is the number of integers k which satisfy $0 < k < n$ and $(k, n) = 1$. This function occurs frequently in the study of elementary number theory; it is called the *Euler Phi Function*. We have just proved in Theorem 5.30 that $\phi(n) = n - 1$ if and only if n is a prime. We close this section by giving another of its properties and, showing in the process, how the ring $\mathbf{Z}_n$ can be used to prove a result about divisibility of integers.

5.32 Euler's Theorem. *If n is a positive integer and a is any nonzero integer with $(a, n) = 1$, then $a^{\phi(n)} - 1$ is divisible by n.*

PROOF. Let $[a_1], \ldots, [a_s]$ be the elements of $U(\mathbf{Z}_n)$; so in particular $\phi(n) = s$. Since $(a, n) = 1$, it follows that $[a]$ is in $U(\mathbf{Z}_n)$. We know that the product of two units is a unit, so the elements $[aa_i] = [a][a_i]$ are also units. Moreover, these are distinct because the equation

$$[a][a_i] = [a][a_j]$$

implies that $[a_i] = [a_j]$, since the $[a]$ can be cancelled. Thus the set

$$[aa_1], \ldots, [aa_s]$$

is a set of s distinct elements of $U(\mathbf{Z}_n)$; this subset must coincide with $U(\mathbf{Z}_n)$. Now let $[p]$ be the product of all the elements in $U(\mathbf{Z}_n)$; we have

$$[p] = [a_1][a_2] \cdots [a_s]$$

and also

$$\begin{aligned} [p] &= [aa_1][aa_2] \cdots [aa_s] \\ &= [a]^s [a_1][a_2] \cdots [a_s] \\ &= [a]^s [p]. \end{aligned}$$

We have used the fact the elements of $U(\mathbf{Z}_n)$ can be expressed in two different ways. Now since $[p]$ is a product of units, $[p]$ is itself a unit and it may be canceled in the last equation to give $[a]^s = [1]$. This is equivalent to the statement that n divides $a^s - 1$, as we were to show.

The case in which n is a prime produces a simpler statement which follows at once from Euler's Theorem.

5.33 Fermat's Little Theorem. *Let p be a prime integer and a any integer not divisible by p. Then $a^{p-1} - 1$ is divisible by p.*

Here are numerical examples of the preceding theorems. The integer $p = 101$ is a prime so Fermat's Theorem implies $2^{100} - 1$ is divisible by 101. To give an illustration in which n is not prime, consider $n = 15$. It is first necessary to determine $\phi(15)$. Make a list of the integers from 1 to 14 and cross out those that are divisible by 3 or 5. Those remaining have g.c.d. with 15 equal to 1; there are eight of these, so $\phi(15) = 8$. (See Exercise 7 below.) Thus $11^8 - 1$ is divisible by 15, since $(11, 15) = 1$.

EXERCISES

1. Find integers x which satisfy the following congruences:

   (i) $3x \equiv 2 \pmod 7$;
   (ii) $12x \equiv 1 \pmod{13}$;
   (iii) $12x \equiv 38 \pmod 5$;
   (iv) $6x \equiv 10 \pmod{14}$.

   In each case interpret the result as an equation over the ring $\mathbf{Z}_n$ for appropriate n.

2. Use the results of Section 4.11 to prove the following statement: There is an integer x which satisfies the congruence $ax \equiv c \pmod n$ if and only if (a, n) is a divisor of c.

3. Compute the product of all the units of $\mathbf{Z}_n$ (as in the proof of 5.32) for the cases $n = 3, 5, 7, 11$.

4. For a prime p, use the facts that $\mathbf{Z}_p$ is a field, that every nonzero element has an inverse, and that at most two elements can be equal to their inverses to prove *Wilson's Theorem*:

   $$(p-1)! \equiv -1 \pmod p.$$

   Here we use the factorial notation $n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$.

5. Find all of the ideals of the ring $\mathbf{Z}_n$ for the cases $n = 4, 6, 8$.

6. Prove that every ideal in $\mathbf{Z}_n$ is principal. Hint: $\mathbf{Z}_n$ is a homomorphic image of $\mathbf{Z}$. Show that the homomorphism maps an ideal of $\mathbf{Z}$ which contains (n) onto an ideal of $\mathbf{Z}_n$ and, moreover, every ideal of $\mathbf{Z}_n$ is obtained this way.

7. Let p and q be prime integers with $p \neq q$. Show $\phi(pq) = (p-1)(q-1) = \phi(p)\phi(q)$.

8. If $[a]$ is the equivalence class relative to congruence modulo n which contains a, prove that if $x, y \in [a]$, then $(x, n) = (y, n)$.

9. Prove that if $(i, n) = 1$, the mapping $\theta: \mathbf{Z}_n \to \mathbf{Z}_n$ defined by $\theta[a] = i[a]$ is a one-one mapping of $\mathbf{Z}_n$ onto $\mathbf{Z}_n$. Is it an isomorphism?

10. An element a of a ring R is said to be *idempotent* if $a^2 = a$. If m and n are relatively prime positive integers greater than 1, prove that the ring $\mathbf{Z}_{mn}$ has at least two idempotents other than the zero and the unity. [Hint: If $1 = mx + ny$, consider $[mx]$ and $[ny]$.]

11. (i) If $n = 2k$, where k is a positive odd integer, verify that $k^2 \equiv k \pmod n$.
    (ii) If n is as in the preceding, prove that the ring $\mathbf{Z}_n$ contains a subring which is isomorphic to $\mathbf{Z}_2$. (A more general result is asked for in Exercise 13 below.)

12. Let m and n be positive integers greater than 1, and let us denote the elements of $\mathbf{Z}_{mn}$ and $\mathbf{Z}_n$ by the coset notation. Verify that the mapping $\theta: \mathbf{Z}_{mn} \to Z_n$ given by $\theta(i + (mn)) = i + (n)$, $i \in \mathbf{Z}$, is well-defined (that is, it actually defines a mapping of $\mathbf{Z}_{mn}$ to $\mathbf{Z}_n$) and that it is in fact a homomorphism of $\mathbf{Z}_{mn}$ onto $\mathbf{Z}_n$.

13. If m and n are relatively prime positive integers greater than 1, prove that the ring $\mathbf{Z}_{mn}$ has a subring which is isomorphic to $\mathbf{Z}_n$. [Hint: If $am + bn = 1$, show that the function $\theta: \mathbf{Z}_n \to \mathbf{Z}_{mn}$ defined by $\theta(i + (n)) = iam + (mn)$, $i \in \mathbf{Z}$, is a well-defined isomorphism of $\mathbf{Z}_n$ with a subring of $\mathbf{Z}_{mn}$. Exercise 10 may help.]

5.5 FINITE FIELDS

For each prime integer p, the ring $\mathbf{Z}_p$ is a field with exactly p elements. We shall prove in this and the next section that a field with only a finite number of elements must contain exactly p^n elements for some prime p and some positive integer n. Moreover, for any such prime power, there is one and only one (up to isomorphism) field with p^n elements.

The Characteristic

Although we are now primarily interested in fields, the concept to be introduced in this section applies to any ring and we therefore give the definition in its general form. We recall that if a is an element of a ring and n is a positive integer, we have given in Section 2.6 a recursive definition of na. We now make the following definition.

> **5.34 Definition.** Let R be a ring. If there exists a positive integer n such that $na = 0$ for every element a of R, the smallest such positive integer n is called the ***characteristic*** of R. If no such positive integer exists, R is said to have ***characteristic zero.***

All the familiar number systems of elementary algebra certainly have characteristic zero. However, let us consider, for example, the ring $\mathbf{Z}_4$ of integers modulo 4. If $[r]$ is any element of this ring, then $2[r] = [r] + [r] = [2r]$ and, generally, if k is a positive integer, $k[r] = [kr]$. The smallest positive integer k such that $[kr] = [0]$ for every element $[r]$ of $\mathbf{Z}_4$ is clearly 4, so $\mathbf{Z}_4$ has characteristic 4. In general, the ring $\mathbf{Z}_n$ has characteristic n.

The definition of the characteristic of a ring makes an assertion about *every* element of the ring. However, in an important special case, the following theorem shows that the characteristic is determined by some one particular element.

> **5.35 Theorem.** *Let R be a ring with a unity e. If there exists a positive integer n such that $ne = 0$, then the smallest such positive integer is the characteristic of R. If no such positive integer exists, then R has characteristic zero.*
>
> PROOF. If n is the smallest positive integer such that $ne = 0$, the characteristic of R certainly cannot be a positive integer less than n. Moreover, if $a \in R$, then
>
> $$na = (na)e = (ne)a = 0a = 0,$$
>
> so that $na = 0$ for every element a of R; hence R has characteristic n. The last sentence of the theorem is an immediate consequence of the definition of characteristic zero.

We know that the ring $\mathbf{Z}_n$ is a field if and only if n is a prime. Hence the characteristic of every field that has been mentioned so far is either zero or a prime. In fact, we shall now prove that this is always true for every integral domain and certainly then for every field.

5.36 Theorem. *Let D be an integral domain with unity e. Define a map θ from $\mathbf{Z}$ into D by the rule $\theta(m) = me$. Then θ is a homomorphism of $\mathbf{Z}$ onto a subring $P = \{me : m \in \mathbf{Z}\}$ of D. If $\ker\theta = (0)$, then D has characteristic zero and P is isomorphic to $\mathbf{Z}$. If $\ker\theta = (n)$ with $n > 0$, then n is prime, n is the characteristic of D, and P is isomorphic to $\mathbf{Z}_n$.*

PROOF. The fact that θ is a homomorphism is an immediate consequence of the definition of multiples me; by definition, θ maps onto the subring P. If $\ker\theta = (0)$, then there is no integer $n \neq 0$ such that $ne = 0 = \theta(n)$. By 5.35, D has characteristic zero. If $\ker\theta = (n)$ with $n > 0$, then n is the least positive integer such that $ne = 0$; for if $me = 0$, then $m \in \ker\theta = (n)$ and so $n \mid m$. Thus D has characteristic n. By the First Isomorphism Theorem 5.27, we have P is isomorphic to $\mathbf{Z}/\ker\theta = \mathbf{Z}_n$.

5.37 Corollary. *If F is a field, then either*

(*i*) *F has characteristic zero and F contains a subfield isomorphic to $\mathbf{Q}$, the rational number field;*

or

(*ii*) *F has characteristic p, p a prime, and F contains a subfield isomorphic to the field $\mathbf{Z}_p$.*

PROOF. Since F is a field, it is an integral domain, and so 5.36 implies the characteristic of F is zero or a prime, p. Moreover, if the characteristic is p, then F contains a subring isomorphic to $\mathbf{Z}_p$. Now if the characteristic is zero, then F contains a subring P isomorphic to $\mathbf{Z}$. If e is the unity of F, then P is the set of multiples me, $m \in \mathbf{Z}$. For such an element me with $m \neq 0$, $(me)^{-1}$ is an element of F. Thus we may define a map from $\mathbf{Q}$ to F by the rule $a/m \to (ae)(me)^{-1}$. We leave it to the reader to prove that this is an isomorphism of $\mathbf{Q}$ with a subfield of F.

The subfield of F which is isomorphic to either $\mathbf{Q}$ or $\mathbf{Z}_p$ is called the *prime subfield* of F.

The literature on the theory of fields is quite extensive. In much of this work the concept of the characteristic of a field plays an essential role. It frequently happens, for example, that although a theorem may be true for every field, different proofs have to be given for the case in which the characteristic is zero and that in which it is a prime.

Extensions of $\mathbf{Z}_p$

Now we study fields which have only a finite number of elements. We shall use the term *order of F* for the number of elements in F. Thus $\mathbf{Z}_p$ is a field of order p.

5.38 Theorem. *Let F be a field with only a finite number of elements. Then F has characteristic p for some prime p,* $\mathbf{Z}_p$ *is isomorphic to a subfield of F, and the order of F is* p^n *for some positive integer n.*

PROOF. Since a finite field cannot contain an infinite subfield isomorphic to $\mathbf{Q}$, 5.37 implies F has characteristic p for some prime p and, moreover, F contains a subfield isomorphic to $\mathbf{Z}_p$. We shall simplify the notation somewhat by assuming that $\mathbf{Z}_p$ is actually a subfield of F. This amounts to identifying $[a]$ with $ae \in F$. Now if $\mathbf{Z}_p = F$, then the number of elements in F is $p = p^1$, and the theorem is true in this case. So we suppose that $\mathbf{Z}_p \neq F$. Then there is an element $c \in F$, $c \notin \mathbf{Z}_p$. The evaluation map

$$\theta_c : \mathbf{Z}_p[x] \to F$$

defined by $\theta_c(f(x)) = f(c)$ maps the infinite ring $\mathbf{Z}_p[x]$ into the finite set F. Such a map cannot be one-to-one so the kernel of θ_c cannot be zero. If the kernel equals $(g(x))$, then by Theorem 4.44, $g(x)$ is an irreducible element of $\mathbf{Z}_p[x]$ which has c as a root. The image of θ_c is a field $\mathbf{Z}_p[c]$ contained in F and every element of $\mathbf{Z}_p[c]$ has a unique expression

$$z = a_0 + a_1c + \cdots + a_{d-1}c^{d-1}$$

if $\deg g(x) = d$. The elements a_i lie in $\mathbf{Z}_p$ and so the number of elements of the form z is p^d. Set $F_1 = Z_p[c]$. If $F_1 = F$, the order of F is p^d, a power of p. If $F_1 \neq F$, then there is an element $s \in F$, $s \notin F_1$. By the same reasoning, the field $F_2 = F_1[s]$ has every element expressible in the unique form

$$w = b_0 + b_1s + \cdots + b_{k-1}s^{k-1},$$

for some k, and each $b_i \in F_1$. So the number of such elements is (order $F_1)^k = p^{dk}$. We may repeat this argument, at each step producing a larger field with order a power of p. Since order of F is finite, the process eventually reaches a step at which $F_t = F$ has order a power of p.

We now know that a finite field has order a power of a prime. The next step in this investigation is to show that for each prime p and each positive integer n, there is a field of order p^n. First we shall record the following very useful result.

5.39 Lemma. *If a and b are elements in a ring of characteristic p, then*

5.40 $$(a+b)^{p^k} = a^{p^k} + b^{p^k},$$

for every positive integer k.

PROOF. First consider the case $k = 1$. The binomial theorem implies

$$(a+b)^p = a^p + C_{p,1}a^{p-1}b + \cdots + C_{p,r}a^{p-r}b^r + \cdots + b^p,$$

where $C_{p,r} = p!/(p-r)!r!$. If $0 < r < p$, then neither term $r!$ nor $(p-r)!$ is divisible by p, and so there is a factor of p in the numerator of $C_{p,r}$. Thus $C_{p,r}$ is an integer which is divisible by p; it follows that $C_{p,r}a^{p-r}b^r = 0$. The only terms remaining are those for $r=0$ and $r=p$. Thus 5.40 holds for $k = 1$. The case for general positive k follows from this case by an easy mathematical induction. We leave this for the reader.

Now we are able to prove the existence result.

5.41 Theorem. *For each prime integer p and each positive integer n, there is a field of order p^n.*

PROOF. Let $h(x)$ be the polynomial in $\mathbf{Z}_p[x]$ defined by

$$h(x) = x^{p^n} - x.$$

By Theorem 5.12, there is a field E containing Z_p and containing elements c_i such that

$$h(x) = (x - c_1)(x - c_2)\cdots(x - c_t), \quad t = p^n.$$

We shall prove that the roots of this polynomial are distinct, so that there are exactly p^n of them, and then prove that this set of roots forms a field.

The root $c = 0$ of $h(x)$ has multiplicity one because $h(x) = x(x^{t-1} - 1)$ and 0 is not a root of the second factor. Now suppose c is any root of $h(x)$ and the multiplicity of c is $m \geq 1$. Then there is a factorization

5.42 $$h(x) = (x - c)^m k(x), \quad k(x) \in E[x].$$

Now consider the polynomial $h(x + c)$ which is obtained from $h(x)$ by substituting $x + c$ in place of x. From the definition of $h(x)$ we have

$$\begin{aligned} h(x + c) &= (x + c)^{p^n} - (x + c) \\ &= x^{p^n} + c^{p^n} - x - c \\ &= h(x) + h(c) \\ &= h(x); \end{aligned}$$

we have used 5.40 to obtain this. Now evaluate $h(x + c)$ by using 5.42 to get

$$h(x) = h(x + c) = x^m k(x + c).$$

Since we saw that the power of x in $h(x)$ was only the first power, it follows that $m = 1$ and so every root has multiplicity one. Thus $h(x)$ has exactly p^n roots in E.

Now we show these roots form a field. To do this we must prove that if a and b are roots of $h(x)$ in E then also $(a + b)$, ab are roots of $h(x)$ and if $a \neq 0$, then a^{-1} is a root. Each of these is straightforward. To show that $a + b$ is a root, simply use 5.40 and the fact that a and b are each roots of $h(x)$. For the product, we have

$$(ab)^{p^n} = a^{p^n} b^{p^n} = ab,$$

where the last equality follows because a and b are roots of $h(x)$. A similar argument shows that a^{-1} is a root whenever a is a nonzero root of $h(x)$. Hence we have proved that the p^n distinct roots of $h(x)$ form a field of order p^n.

A drawback to the above proof is that it does not give a direct construction of a field of order p^n. Here is a method which gives the construction provided we are able to find a suitable polynomial. If $g(x)$ is a polynomial in $\mathbf{Z}_p[x]$, and $g(x)$ is irreducible of degree n, then $\mathbf{Z}_p[x]/g(x)$ is a field by Theorem 5.10 and the number of elements in this field is p^n because every coset $u(x) + (g(x))$ has a unique representative which has the form $a_0 + a_1x + \cdots + a_{n-1}x^{n-1}$, with the $a_i \in \mathbf{Z}_p$. Since there are p choices for each a_i, there are p^n elements in all. It has not been proved that such a polynomial always exists. Although this is a true statement, its proof will not be given until some of the theory of groups is presented in later chapters. The field obtained by this method—namely, as $\mathbf{Z}_p[x]/(g(x))$ for an irreducible polynomial $g(x)$ having degree n, is called the *Galois Field of order* p^n and is sometimes denoted as $GF(p^n)$. It will be seen later that any two fields of the same order are isomorphic, so it is appropriate to call this *the* field of order p^n.

For now we give some examples where the required polynomial is easily found. A field of order four is constructed by first finding an irreducible polynomial of degree two in $\mathbf{Z}_2[x]$. The total number of polynomials of degree 2 is 4; the reducible ones are x^2, $x(x + 1)$, $(x + 1)^2$. The remaining one is

$x^2 + x + 1 = g(x)$ and it is irreducible. In the field $\mathbf{Z}_2[x]/g(x)$, we let c denote the coset $x + (g(x))$. Then since c is a root of $g(x)$, we have $c^2 = c + 1$. The four elements of the field are $0, 1, c, 1 + c$. The full addition and multiplication tables are given below.

$+$	0	1	c	$1+c$
0	0	1	c	$1+c$
1	1	0	$1+c$	c
c	c	$1+c$	0	1
$1+c$	$1+c$	c	1	0

	0	1	c	$1+c$
0	0	0	0	0
1	0	1	c	$1+c$
c	0	c	$1+c$	1
$1+c$	0	$1+c$	1	c

Addition and Multiplication Tables
for the Field of Order Four

In order to construct a field of order 8 by this method, it is necessary to find an irreducible polynomial in $\mathbf{Z}_2[x]$ having degree 3. It turns out that there are two such polynomials: $g(x) = x^3 + x + 1$ and $f(x) = x^3 + x^2 + 1$. Let c denote the coset $x + (g(x))$ in $\mathbf{Z}_2[x]/g(x)$. Every element in this field has the form $a_0 + a_1c + a_2c^2$, with the $a_i \in \mathbf{Z}_2$. We shall not write out the full table for this field, but we shall make a few representative computations. The powers of c are found using the basic relation $c^3 = c + 1$, which follows from $g(c) = 0$. The powers beyond the third are

$$\begin{aligned} c^4 &= cc^3 = c(1 + c) = c + c^2; \\ c^5 &= cc^4 = c(c + c^2) = c^2 + c^3 = c^2 + c + 1; \\ c^6 &= (c^3)^2 = (c + 1)^2 = c^2 + 1; \\ c^7 &= cc^6 = c(c^2 + 1) = c^3 + c = (1 + c) + c = 1. \end{aligned}$$

It may be of interest to note that $c + 1$ is a root of the other irreducible polynomial of degree three; that is, $f(c + 1) = 0$.

EXERCISES

1. Find the factorization of the given polynomials as a product of irreducible polynomials in the indicated $\mathbf{Z}_p[x]$:

   (a) $x^9 + 1$ in $\mathbf{Z}_3[x]$;
   (b) $x^3 + x^2 + x + 1$ in $\mathbf{Z}_2[x]$;
   (c) $x^6 - x^3 + 1$ in $\mathbf{Z}_3[x]$;
   (d) $x^{2p} - (2x)^p + 1$ in $\mathbf{Z}_p[x]$.

2. By testing each element of $\mathbf{Z}_p$, find all the roots in $\mathbf{Z}_p$ and their multiplicities for polynomials and the given values of p:

   (a) $x^4 + x^3 + x + 1, p = 2, p = 3$;
   (b) $x^5 + x^4 - x - 1, p = 3, p = 5$.

3. The polynomial $x^2 + 1$ is prime in $\mathbf{Z}_3[x]$, so $F = \mathbf{Z}_3[x]/(x^2 + 1)$ is a field of order 9. Let c be a root of $x^2 + 1$. Find the irreducible polynomials in $\mathbf{Z}_3[x]$ which have $c + 1$ and $c - 1$ as roots.

4. Continuing with the notation of the preceding problem, express the powers of $c + 1$ in the form $a + bc$, with $a, b \in \mathbf{Z}_3$.

5. Prove that there are $(p^2 - p)/2$ monic quadratic polynomials which are prime over the field $\mathbf{Z}_p$. [Hint: First determine the number that are *not* prime.]

6. Prove that there are $(p^3 - p)/3$ monic cubic polynomials which are prime over the field $\mathbf{Z}_p$.

5.6 GALOIS THEORY OF FINITE FIELDS

Galois Theory is the study of the extensions of a given field K. The aim of the study is to classify in some way all the fields, subject to some restrictions, which contain K. The main result in this subject takes an extension field F of K, subject to certain assumptions, and gives a description of all the subfields of F which contain K in terms of an associated, presumably simpler, object called the Galois Group of F over K. We shall not give definitions of these terms here for that would require considerable preparation. We shall only illustrate this important theory by giving a classification of all the subfields of a finite field of order p^n. The role of K will be played by $\mathbf{Z}_p$. In this case the associated "simpler object" will be the positive integer n. The classification theorem will say, in part, that if F is a field of order p^n, then the subfields of F correspond in a one-to-one way with the divisors of the integer n. This will imply, for example, that a field with order p^{12} has exactly six subfields; they are the fields with orders p, p^2, p^4, p^3, p^6, and p^{12}.

We begin by showing that, up to isomorphism, there is only one finite field having a given order. A preparatory lemma similar to Euler's Theorem 5.32 is needed.

5.43 Lemma. *If a field F has order k, then every element of F is a root of $x^k - x$.*

PROOF. Clearly the zero element is a root of the polynomial. Suppose a is a nonzero element of F. Let $b_1, b_2, \ldots, b_{k-1}$ be the nonzero elements of F and let u be their product. Then $ab_1, ab_2, \ldots, ab_{k-1}$ is also the set of nonzero elements of F so

$$u = (ab_1)(ab_2)\cdots(ab_{k-1}) = a^{k-1}b_1\cdots b_{k-1} = a^{k-1}u.$$

Since u is a unit it follows that $a^{k-1} = 1$ and $a^k = a$.

5.44 Theorem. *If F and L are fields of order p^n, then F is isomorphic to L.*

PROOF. Since F and L each have order a power of p, each has characteristic p and each contains a subfield isomorphic to $\mathbf{Z}_p$, by Theorem 5.38. Let P be a subfield of F and let R be a subfield of L; suppose that there is an isomorphism $\theta: P \to R$. We know that such an isomorphism exists between the prime subfields of F and L since each is isomorphic to $\mathbf{Z}_p$. We take a more general point of view and do not necessarily assume that P and R are the prime subfields but any subfields whatever which are isomorphic. We shall prove the following:

5.45 If $P \neq F$, then there is a subfield P_1 of F which is larger than P and which is isomorphic to a subfield R_1 of L.

If this is assumed for a moment, then we may complete the proof by repetition of the same argument. That is, if $P_1 \neq F$, then there is a subfield P_2 of F which is larger than P_1 and which is isomorphic to a subfield of L. Since F is finite, the chain of subfields $P < P_1 < P_2 < \cdots$ must eventually reach a subfield P_t which equals F and is isomorphic to a subfield of L. Since the order of L equals the order of F, the isomorphism must carry F onto L.

So we turn to the proof of the assertion 5.45. The results from Section 4.7 will be used repeatedly in this proof. Since $P < F$, there is an element $c \in F$ with $c \notin P$. The evaluation map $\theta_c: P[x] \to F$ maps the infinite ring $P[x]$ into the finite ring F; the map cannot be one-to-one. Hence $\ker\theta_c = (g(x)) \neq (0)$, for some irreducible polynomial $g(x)$ in $P[x]$, and $P_1 = P[c]$ is a subfield of F which is larger than P because $c \in P_1$, but $c \notin P$. Now we must select the proper element c' in L so that we may conclude $P[c]$ is isomorphic to $R[c']$. The selection of c' is made by using the isomorphism between P and R to obtain an isomorphism between $P[x]$ and $R[x]$. If ϕ is the isomorphism from P onto R, then the map (which we also call ϕ) defined by

$$\textbf{5.46} \quad \phi(a_0 + a_1x + \cdots + a_mx^m) = \phi(a_0) + \phi(a_1)x + \cdots + \phi(a_m)x^m$$

is an isomorphism from $P[x]$ to $R[x]$. There are several details left to the reader for this point (see Exercise 3 following Section 5.3). Let $G(x) = \phi(g(x))$; so $G(x)$ is an irreducible polynomial in $R[x]$. We shall select c'

to be an element of L which is a root of $G(x)$. It is necessary to show that $G(x)$ has a root in L. To see this observe first that c is a root of the polynomial

$$h(x) = x^{p^n} - x$$

by 5.43 and, since $g(x)$ is the irreducible polynomial in $P[x]$ having c as a root, it follows that

$$h(x) = g(x)k(x), \quad k(x) \in P[x].$$

Now apply ϕ to this equation to get

$$\phi(h(x)) = \phi(g(x))\phi(k(x)) = G(x)K(x).$$

Since the coefficients of $h(x)$ are just ± 1, and since we are writing 1 for the unity element of both F and L, it follows that $\phi(h(x)) = h(x)$; thus

$$x^{p^n} - x = G(x)K(x).$$

Now by Theorem 5.43, every root of the polynomial on the left side of this equation is an element of L. Thus $G(x)$ has a root in L which we call c'. The isomorphism from $P[c]$ to $R[c']$ is obtained as follows. We define a mapping

$$\alpha: P[x]/(g(x)) \rightarrow R[x]/(G(x))$$

by the rule

$$\alpha(u(x) + (g(x))) = \phi(u(x)) + (G(x)).$$

Since ϕ is an isomorphism of $P[x]$ onto $R[x]$ which carries $g(x)$ to $G(x)$, one can prove that α is an isomorphism of the fields. Now $P[c]$ is isomorphic to $P[x]/(g(x))$ and $R[c']$ is isomorphic to $R[x]/(G(x))$, so it follows that $P[c]$ is isomorphic to $R[c']$, and we have finally finished the proof.

Here is an immediate corollary of the theorem.

5.47 Corollary. *If F is a field of order p^n, then F is the splitting field of the polynomial $x^{p^n} - x$.*

PROOF. In the proof of Theorem 5.41, we proved the existence of a field of order p^n which was a splitting field for the above polynomial. Now we know that any field of order p^n must be isomorphic to the field constructed there, so the result follows.

Before giving the classification theorem, we shall require the following.

5.48 Theorem. *Let k and m be positive integers and let F be any field. Then $x^k - 1$ divides $x^m - 1$ in $F[x]$ if and only if k divides m.*

PROOF. Let us apply the division algorithm to obtain $m = kq + r$, with $0 \le r < k$. Also let

$$h_{m,k}(x) = x^{m-k} + x^{m-2k} + \cdots + x^{m-qk}.$$

Then the division algorithm in $F[x]$ gives us

$$x^m - 1 = (x^k - 1)h_{m,k}(x) + (x^r - 1).$$

Thus the remainder after division of $x^m - 1$ by $x^k - 1$ is zero if and only if the remainder after division of m by k is zero.

This idea can be used to prove a result about the divisibility of integers in place of polynomials.

5.49 Corollary. *Let m and k be positive integers and let p be any integer with $p \ge 2$. Then $p^k - 1$ divides $p^m - 1$ if and only if k divides m.*

PROOF. To prove this, use the proof of the Theorem 5.48 and replace x by p. The equation in the division algorithm remains valid and the proof carries over to this case without further changes.

Finally we have all the necessary tools at our disposal to prove the main result of the section.

5.50 Theorem. *Let F be a field of order p^n. For each divisor d on n, there is one and only one subfield of F which has order p^d. It equals the set F_d of all the roots of the polynomial $x^{p^d} - x$. If a subfield N of F has order p^t, then $t \mid n$ and $N = F_t$.*

PROOF. If N is any subfield of F (subfield will always mean a nonzero subfield), then N has the same characteristic as F and so the order of N is p^t for some positive integer t. It will be convenient in the rest of the proof to use polynomials $H_k(x)$ and $G_k(x)$ defined for any positive integer k by

$$H_k(x) = x^{p^k} - x = xG_k(x).$$

If N is a subfield of F having order p^t, then every root of $H_t(x)$ is an element of N. Every element of F (in particular, every element of N) is also a root of $H_n(x)$. It follows that $H_t(x)$ divides $H_n(x)$; after cancelling a factor of x, it follows that $G_t(x)$ divides $G_n(x)$. Now apply Theorem 5.48 to conclude $p^t - 1$ divides $p^n - 1$; and by Corollary 5.49, t divides n. So we have proved that the order of a subfield of F must have order p^t with t a divisor of n. Now let t be any divisor of n. Then $H_t(x)$ divides $H_n(x)$, and since $H_n(x)$ has p^n distinct roots in F, exactly p^t of these elements are roots of $H_t(x)$. We have already seen in the proof of Theorem 5.41 that this set of p^t elements forms a subfield of F. Hence to each divisor t of n there corresponds the subfield F_t of order p^t. The last assertion in the theorem which we must prove is that there is just one subfield of order p^t for a divisor t of n. If M and N are subfields of F, each having order p^t, then every element of M and every element of N is a root of $H_t(x)$. Since $H_t(x)$ has degree p^t, there can be no more than p^t roots of this polynomial in F. It follows that $M = N$, and the proof of the theorem is complete.

Let us give a few examples of how the theorems proved in the last few sections may be used to draw some conclusions about polynomial equations over finite fields.

5.51 If $g(x)$ is an irreducible polynomial of degree d in $\mathbf{Z}_p[x]$, then $g(x)$ can be written as a product of factors of degree one in $F[x]$ whenever the order of F is p^n and $d \mid n$.

This follows because $N = \mathbf{Z}_p[x]/g(x)$ is a field of order p^d and contains a root of $g(x)$. If F has order p^n with $d \mid n$, then, by Theorem 5.48, F must contain an isomorphic copy of N and so F contains a root of $g(x)$. This implies $g(x)$ divides $H_n(x)$ and hence $g(x)$ is a product of degree one factors in $F[x]$, because this is so for $H_n(x)$.

5.52 If $g(x)$ is a polynomial of positive degree in $\mathbf{Z}_p[x]$, and if c is an element of some extension field of $\mathbf{Z}_p$ which is a root of $g(x)$, then c^p is also a root of $g(x)$.

If $g(x) = a_0 + a_1x + \cdots + a_nx^n$, with the $a_i \in \mathbf{Z}_p$, then

$$\begin{aligned} g(c^p) &= a_0 + a_1c^p + \cdots + a_nc^{pn} \\ &= (a_0 + a_1c + \cdots + a_nc^n)^p \\ &= g(c)^p = 0. \end{aligned}$$

In these equations we have used 5.40 and Fermat's Little Theorem which asserts that $a^p = a$ for each a in $\mathbf{Z}_p$. This result can be of help in factoring polynomials.

As an example, consider the polynomial $x^2 + 1$ in $\mathbf{Z}_3(x)$. It is an irreducible polynomial. If c is one of its roots in an extension field, then

c^3 must be another. Moreover $c^3 \neq c$, since $c^2 = -1 \neq 1$. Thus we have a factorization

$$x^2 + 1 = (x + c)(x + c^3)$$

over the extension field containing c.

EXERCISES

In these exercises, p is a prime integer and $H_n(x)$ is the polynomial in $\mathbf{Z}_p[x]$ defined by $H_n(x) = x^{p^n} - x$.

1. If $g(x)$ is an irreducible polynomial of degree d in $\mathbf{Z}_p[x]$, then $g(x)$ divides $H_n(x)$ if and only if $d \mid n$.
2. Prove $H_n(x)$ is the product of every monic, irreducible polynomial in $\mathbf{Z}_p[x]$ which has degree dividing n.
3. If n is a prime integer, then $\mathbf{Z}_p[x]$ has $(p^n - p)/n$ monic, irreducible polynomials of degree n. [Hint: Examine the factors of $H_n(x)$ by using the previous exercises. This same idea can be applied to count the number of irreducible polynomials of degree n when n is a prime power. A much more subtle counting method can be used to show there is always an irreducible polynomial of degree n for any n.]
4. Let $g(x)$ be an irreducible polynomial of degree 4 in $\mathbf{Z}_p[x]$ and let F be a splitting field for $g(x)$ of order p^4. Let L be the subfield of F having order p^2. Prove that $g(x)$ is the product of two irreducible factors of degree 2 in $L[x]$. [Hint: To show that $g(x)$ is not irreducible over L, consider the possibility that it is irreducible and examine the field $L[x]/g(x)$.]
5. Let $g(x)$ be an irreducible polynomial of degree 2 in $\mathbf{Z}_p[x]$ and let F be the field of order p^3. Show that $g(x)$ is irreducible in $F[x]$.

5.7 PRIME POLYNOMIALS OVER THE RATIONAL FIELD

In this section we apply some information about polynomials over the finite fields $\mathbf{Z}_p$ to obtain information about polynomials over the rational field Q. We seek information that will aid in determining if a given polynomial is prime. Generally this is a very difficult problem. Of course, for polynomials of small degree, this will not be a serious problem. Every polynomial of degree one is prime. For polynomials of degree two or three the Theorem 4.65 will help us decide because a polynomial of degree two or three is prime if and only if it has no root in Q. However, even for polynomials of degree four, we may encounter some difficulties. Such a polynomial might not have a rational root but may be the product of two irreducible polynomials of degree two. The results in this section will be rather special in that fairly restrictive conditions must hold before conclusions may be drawn.

We shall prove two theorems about polynomials with rational coefficients. As usual, $\mathbf{Z}[x]$ denotes the ring of polynomials with integer coefficients. It will be seen that questions about rational polynomials may often be answered by using facts about polynomials in $\mathbf{Z}[x]$.

5.53 Lemma. *Let $f(x)$, $g(x)$, and $h(x)$ be elements of the ring $\mathbf{Z}[x]$ such that $f(x) = g(x)h(x)$. If p is a prime integer which is a divisor of every coefficient of $f(x)$, then p is a divisor of every coefficient of $g(x)$ or a divisor of every coefficient of $h(x)$.*

PROOF. Let us define a homomorphism $\psi:\mathbf{Z}[x] \to \mathbf{Z}_p[x]$ by the rule

$$\psi(a + bx + \cdots + gx^n) = a^* + b^*x + \cdots + g^*x^n \tag{5.54}$$

where for any integer a, a^* denotes the coset $a + (p)$ in $\mathbf{Z}_p$. The verification that ψ is a homomorphism is straightforward. Now suppose that the polynomial $f(x) \in \mathbf{Z}[x]$ has the factorization $f(x) = g(x)h(x)$ with $g(x)$ and $h(x)$ also in $\mathbf{Z}[x]$. The assumption that p divides every coefficient of $f(x)$ is equivalent to the statement that $\psi(f(x)) = 0$. It follows that $\psi(g(x))\psi(h(x)) = 0$. Since $\mathbf{Z}_p[x]$ is an integral domain either $\psi(g(x)) = 0$ or $\psi(h(x)) = 0$. This implies p divides every coefficient of $g(x)$ or p divides every coefficient of $h(x)$, as we wished to prove.

The following lemma, whose proof will be based on the preceding lemma, shows that a polynomial with *integral* coefficients is prime over the field $\mathbf{Q}$ if and only if it cannot be factored into a product of two polynomials of positive degree with *integral* coefficients. It will then be possible to prove that certain polynomials are prime over $\mathbf{Q}$ by making use of special properties of the integers.

5.55 Lemma. *Let $f(x)$ be an element of $\mathbf{Z}[x]$ such that $f(x) = g(x)h(x)$, where $g(x)$, $h(x) \in \mathbf{Q}[x]$. Then there exist polynomials $g'(x)$, $h'(x)$ of $\mathbf{Z}[x]$ having the same degrees as $g(x)$ and $h(x)$, respectively, such that $f(x) = g'(x)h'(x)$.*

PROOF. Let k be the l.c.m. of the denominators of the coefficients of $g(x)$, so that $kg(x)$ has integral coefficients. Similarly, let l be an integer such that $lh(x)$ has integral coefficients. Since $f(x) = g(x)h(x)$, it follows that

$$klf(x) = g_1(x)h_1(x), \tag{5.56}$$

where $g_1(x)$ and $h_1(x)$ have integral coefficients. We may then apply the preceding lemma as follows. If p is a prime divisor of kl, it must be a divisor of all coefficients of $g_1(x)$ or of $h_1(x)$; hence p can be divided from both sides of Equation 5.56, and we still have polynomials with integral coefficients. By a repetition of this process, we can divide out every prime factor of kl and finally get $f(x) = g'(x)h'(x)$, where $g'(x)$ and $h'(x)$ have integral coefficients. It is almost trivial that $g'(x)$ has the same degree as $g(x)$, and also that $h'(x)$ has the same degree as $h(x)$. The proof is therefore complete.

We are now ready to prove the following theorem of Eisenstein.

5.57 Theorem. *Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ be a polynomial of positive degree n over the ring* **Z** *of integers, and p a prime integer such that $a_i \equiv 0 \pmod p$ for $i = 0, 1, \ldots, n-1$; $a_n \not\equiv 0 \pmod p$ and $a_0 \not\equiv 0 \pmod{p^2}$. Then $f(x)$ is a prime polynomial over* **Q**.

PROOF. The preceding lemma shows that we need only prove that $f(x)$ cannot be factored into a product of two factors of positive degree over **Z**. Let us assume that

5.58
$$\begin{aligned} a_0 &+ a_1x + \cdots + a_nx^n \\ &= (b_0 + b_1x + \cdots + b_mx^m)(c_0 + c_1x + \cdots + c_kx^k), \\ &= g(x)h(x) \end{aligned}$$

with $\deg g(x) = m$ and $\deg h(x) = k$. Here all the coefficients are integers and $m + k = n$. Let $\psi\colon \mathbf{Z}[x] \to \mathbf{Z}_p[x]$ be the homomorphism defined by 5.54 which reduces the coefficients modulo p. By assumption on the coefficients of $f(x)$ we have $\psi(f(x)) = a_n^*x^n$. Since this equals $\psi(g(x))\psi(h(x))$, it follows that

$$\psi(g(x)) = b_m^*x^m \quad \text{and} \quad \psi(h(x)) = c_k^*x^k$$

since each of these divides $a_n^*x^n$. We make the supposition that the factorization of $f(x)$ is nontrivial; that is, that both m and k are at least one. Then the coefficients b_j for $j < m$ are all congruent to zero modulo p and, in particular, this means $b_0 \equiv 0 \pmod p$. Similarly, $c_0 \equiv 0 \pmod p$. But then $a_0 = b_0c_0$ is divisible by p^2, contrary to the assumptions made in the hypothesis. Thus no such factorization of $f(x)$ can exist.

5.59 Corollary. *If n is an arbitrary positive integer, there exist polynomials of degree n over* **Q** *that are prime over* **Q**.

PROOF. This result is easily established by examples. As an illustration, the polynomial $x^n - 2$ over **Q** satisfies all the conditions of the preceding theorem with $p = 2$. Hence $x^n - 2$ is a prime polynomial over Q for each positive integer n. In like manner, each of the following polynomials of degree n over Q is prime over Q: $x^n + 2$, $x^n + 3$, $3x^n + 2x^{n-1} + 2x^{n-2} + \cdots + 2x + 2$, $x^n + 9x + 3$ $(n > 1)$. The reader will have no difficulty in constructing other examples.

The homomorphism $\psi: \mathbf{Z}[x] \to \mathbf{Z}_p[x]$ which reduces the coefficients modulo p can occasionally be useful for directly proving that a polynomial is prime, as we now show.

5.60 Theorem. *Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ be a polynomial in $\mathbf{Z}[x]$ and let p be a prime integer which does not divide the leading coefficient a_n of $f(x)$. If $\psi(f(x))$ is a prime polynomial in $\mathbf{Z}_p[x]$, then $f(x)$ is a prime polynomial in $\mathbf{Q}[x]$.*

PROOF. If $f(x)$ is not prime in $\mathbf{Q}[x]$, then it has a factorization $f(x) = g(x)h(x)$ with $g(x)$ and $h(x)$ in $\mathbf{Z}[x]$ and $\deg g(x) = m$, $\deg h(x) = k$, $m > 0$ and $k > 0$. The leading coefficient of $f(x)$, namely a_n, is not divisible by p but is equal to the product of the leading coefficients of $g(x)$ and $h(x)$, respectively. Thus neither of these leading coefficients is divisible by p and it follows that $\deg \psi(g(x)) = \deg g(x)$ and $\deg \psi(h(x)) = \deg h(x)$. But $\psi(g(x))\psi(h(x)) = \psi(f(x))$ is a prime polynomial and cannot be factored as a product of two polynomials having positive degree. This is an impossible situation caused by the assumption that $f(x)$ was not prime in $\mathbf{Q}[x]$. Thus $f(x)$ is prime after all.

As a very simple application of this result we use the fact that $x^3 + x + 1$ is a prime polynomial in $\mathbf{Z}_2[x]$ and conclude that every polynomial of the form $x^3 + 2ax^2 + (2k+1)x + (2t+1)$ is prime in $\mathbf{Q}[x]$, for every choice of integers a, k, t. For a specific example we may assert that $x^3 + 4x^2 + 9x + 225$ is prime. It would be much more work to verify this by testing all the possible rational roots since 18 roots would have to be tested.

EXERCISES

1. Show that each polynomial is prime in $\mathbf{Q}[x]$:

(a) $x^3 + 4x^2 - 3x + 5$,
(b) $x^4 - 6x^2 + 3x - 12$,
(c) $x^4 + 4x^2 - 3x + 5$,
(d) $22x^3 + 5x + 35$.

2. Apply Eisenstein's Theorem to $f(x+1)$ to prove that $f(x) = 1 + x + \cdots + x^{p-1} = (x^p - 1)/(x - 1)$ is prime in $\mathbf{Q}[x]$ for any prime number p.
3. State and prove an analogue of Eisenstein's Theorem which applies to polynomials in $F[x]$, for any field F which is the field of quotients of a Euclidean domain R.
4. Use the version of Eisenstein's Theorem implied in the previous exercise to show that $x^4 + 4x^3 + 3(1 + i)x + (1 + i)$ is prime in $F[x]$, where $F = \mathbf{Q}[i]$, and R is the ring of Gaussian integers $\mathbf{Z}[i]$.

5.8 POLYNOMIALS OVER THE REAL OR COMPLEX NUMBERS

In this section we shall discuss some properties of polynomials over the field **R** of real numbers or the field **C** of complex numbers. We begin with a few remarks, essentially established in elementary algebra, about quadratic polynomials, that is, polynomials of degree 2.

Let

$$g(x) = ax^2 + bx + c, \qquad a \neq 0,$$

be a quadratic polynomial with coefficients in the field **C**. Then it is well known that the polynomial $g(x)$ has roots r_1 and r_2, where

5.61
$$r_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad r_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

We may point out that, by a special case of Theorem 3.55, every nonzero complex number has two square roots. Hence r_1 and r_2, given by 5.61, are complex numbers and it is easy to verify by direct calculation that

5.62
$$g(x) = a(x - r_1)(x - r_2).$$

Since these first-degree factors have coefficients in **C**, it is apparent that no quadratic polynomial over **C** is a prime polynomial over **C**.

It is customary to call $b^2 - 4ac$ the *discriminant* of the quadratic polynomial $ax^2 + bx + c$. For convenience, let us designate this discriminant by D.

From 5.61 it follows that $r_1 = r_2$ if and only if $D = 0$. However, the factorization 5.62 holds in any case, so $D = 0$ is a necessary and sufficient condition that the polynomial $g(x)$ have a double root.

Now let us assume that the quadratic polynomial $g(x)$ has *real* coefficients. Then the roots r_1 and r_2 will also be real if and only if $D \geq 0$, for only in this case will D have real square roots. The factorization 5.62 of $g(x)$ into factors of the first degree is therefore a factorization over **R** if and only if $D \geq 0$. If $D < 0$, $g(x)$ has no real root and $g(x)$ is therefore prime over **R**.

Let us summarize some of these observations in the following theorem.

5.63 Theorem. *No quadratic polynomial over the field* **C** *of complex numbers is prime over* **C**. *A quadratic polynomial over the field* **R** *of real numbers is prime over* **R** *if and only if its discriminant is negative.*

We have referred above to Theorem 3.55, where it was proved by use of the trigonometric form of a complex number that every nonzero complex number has n nth roots. It may be worth pointing out that the *square* roots of a complex number may also be computed by an algebraic process. As an illustration, let us seek the roots of the polynomial $x^2 + x - (1 + 3i)$ over **C**. By 5.61, these roots can immediately be written down in the form

$$\frac{-1 \pm \sqrt{5 + 12i}}{2} \tag{5.64}$$

Now in order to express these roots in the usual form of complex numbers, we need to compute the square roots of $5 + 12i$. To do so, suppose that s and t are unknown real numbers such that $s + ti$ is a square root of $5 + 12i$. Thus we have

$$(s + ti)^2 = 5 + 12i,$$

or

$$s^2 - t^2 + 2sti = 5 + 12i.$$

In turn, this implies both of the following equations involving the real numbers s and t:

$$s^2 - t^2 = 5, \quad 2st = 12.$$

If we solve these two simultaneous equations by elementary methods and remember that s and t are real (so that $s^2 \geq 0$ and $t^2 \geq 0$), we find the solutions to be $s = 3$, $t = 2$ and $s = -3$, $t = -2$. Hence, the square roots of $5 + 12i$ are $\pm(3 + 2i)$. Substituting in 5.64, we find that the roots of the polynomial $x^2 + x - (1 + 3i)$ are $1 + i$ and $-(2 + i)$.

We have shown above that no quadratic polynomial is prime over **C**. Another special case of some interest is the following. Let us consider a polynomial of the form $ax^n + b$, where a and b are nonzero complex numbers and n is an arbitrary positive integer greater than 1. Since, by Theorem 3.55, the complex number $-b/a$ has n distinct nth roots and these are obviously roots of the polynomial $ax^n + b$, this polynomial can be factored over **C** into a product of factors of the first degree. In particular, such a polynomial can never be prime over **C**.

The general theorem which we shall next state is partially suggested by the special cases already discussed. The theorem was first proved by the famous German mathematician Carl Friedrich Gauss (1777–1855), and is of such importance that it has often been called "The Fundamental Theorem of Algebra." There is no elementary proof of this theorem. In the next section we shall give a proof that is perhaps as elementary a proof as is possible. It is based upon properties of functions of two real variables that are usually proved in the study of advanced calculus.

5.65 Theorem. *If $f(x)$ is an element of* $\mathbf{C}[x]$ *of positive degree, there exists an element of* $\mathbf{C}$ *which is a root of the polynomial* $f(x)$.

If r is a complex number which is a root of the polynomial $f(x)$ of degree n over $\mathbf{C}$, then in $\mathbf{C}[x]$ we can use the Factor Theorem and write

$$f(x) = (x - r)f_1(x),$$

where $f_1(x)$ is of degree $n - 1$. It is then apparent from this observation and Theorem 4.39 that the preceding theorem can be expressed in either of the following alternate forms.

5.66 Theorem. *The only prime polynomials of* $\mathbf{C}[x]$ *are the polynomials of the first degree.*▪

5.67 Theorem. *If $f(x)$ is an element of* $\mathbf{C}[x]$ *of positive degree, then $f(x)$ is itself of the first degree or it can be factored in* $\mathbf{C}[x]$ *into a product of polynomials of the first degree.*

We next consider the question of which polynomials over the real field $\mathbf{R}$ are prime over $\mathbf{R}$. Of course, the polynomials of the first degree are always prime, and we have shown in Theorem 5.63 that the quadratic polynomials over $\mathbf{R}$ that are prime over $\mathbf{R}$ are those with negative discriminant. A little later we shall prove that these are the only prime polynomials over $\mathbf{R}$. First, however, we need a preliminary result, which is of some interest in itself.

Let

$$f(x) = a_nx^n + \cdots + a_1x + a_0$$

be a polynomial of positive degree with real coefficients. Since $\mathbf{R} \subset \mathbf{C}$, then also $f(x) \in \mathbf{C}[x]$ and Theorem 5.65 states that there exists an element r of $\mathbf{C}$ such that $f(r) = 0$. We now want to make use of the concept of the conjugate of a complex number, introduced in Section 3.10. We recall that if $u = a + bi$ is a complex number, then the conjugate u^* of u is defined by $u^* = a - bi$. It was shown that the mapping $u \to u^*$ is a one-one mapping of **C** onto **C**, which preserves the operations of addition and multiplication. Now since $f(x)$ is assumed to have real coefficients and a real number is equal to its conjugate, it is not difficult to verify that

$$[f(r)]^* = a_n(r^*)^n + \cdots + a_1r^* + a_0 = f(r^*).$$

But, since $f(r) = 0$, it follows that $[f(r)]^* = 0$ and therefore $f(r^*) = 0$. That is, r^* is also a root of the polynomial $f(x)$. This result we state as the following theorem.

5.68 Theorem. *If r is a complex number which is a root of the polynomial $f(x)$ with real coefficients, then the conjugate r^* of r is also a root of $f(x)$.*

If it happens that r is a real number, then $r^* = r$, and this theorem has no content. However, if r is not real, then r^* and r are distinct roots of $f(x)$. It follows that in $\mathbf{C}[x]$ we have

$$f(x) = (x - r)(x - r^*)f_1(x),$$

with the degree of $f_1(x)$ two less than the degree of $f(x)$. If $r = a + bi$, then $r^* = a - bi$, and a simple calculation shows that

$$(x - r)(x - r^*) = x^2 - 2ax + a^2 + b^2.$$

We can therefore write

5.69
$$f(x) = (x^2 - 2ax + a^2 + b^2)f_1(x),$$

and the quadratic factor on the right clearly has *real* coefficients. Since also $f(x)$ has real coefficients, it is easy to verify that $f_1(x)$ must have real coefficients. It follows that 5.69 gives a factorization of $f(x)$ in $\mathbf{R}[x]$. If $\deg f(x) > 2$, $f(x)$ can therefore not be a prime polynomial over **R**. This result, combined with Theorem 5.63, completes the proof of the following theorem.

5.70 Theorem. *The only polynomials of $\mathbf{R}[x]$ that are prime over **R** are the polynomials of the first degree and the quadratic polynomials with negative discriminant.*

Now a polynomial of *odd* degree clearly cannot be factored into a product of quadratic polynomials. Therefore, if a polynomial $f(x)$ of $\mathbf{R}[x]$ of odd degree is expressed as a product of prime polynomials over **R**, at least one of these prime polynomials (in fact, an odd number of them) must be of the first degree. This implies that $f(x)$ has at least one real root, and the following is therefore an almost immediate consequent of the preceding theorem.

5.71 Corollary. *A polynomial with real coefficients and of odd degree necessarily has a real root.*

Except for quadratic polynomials, and polynomials of the special form $ax^n + b$, we have not given any indication as to how one might actually *find* the real or complex roots of a given polynomial. This is a difficult problem but some information can be found in texts on the "theory of equations." In particular, there are algebraic formulas for the roots of polynomials of degrees 3 or 4 with real or complex coefficients. Although these formulas are of great theoretical interest, they are not convenient to use in a numerical case. It is, however, not too difficult to develop methods of approximating the roots to any desired accuracy, and this is what is usually done in practical applications.

EXERCISES

1. Find the roots of each of the following polynomials and express each root in the standard form $a + bi$ of a complex number:

   (a) $x^2 - (3i - 2)x - 5 - i$,
   (b) $x^2 + ix + 1$,
   (c) $x^2 - (2 + i)x - 1 + 7i$,
   (d) $x^2 + x + 4$,
   (e) $x^2 - x + 2 + \sqrt{2}i$,
   (f) $x^2 + 2x + i$.

2. Factor each of the following polynomials of $\mathbf{R}[x]$ into a product of prime polynomials over $\mathbf{R}$:

   (a) $x^3 - 2x - 4$,
   (b) $x^3 - x^2 - 3x + 6$,
   (c) $x^4 + 1$,
   (d) $x^4 + 2x^2 - 8$,
   (e) $x^4 + x^3 + 2x^2 + x + 1$,
   (f) $x^5 + 1$.

*5.9 THE FUNDAMENTAL THEOREM OF ALGEBRA

In this section we shall give a proof of Theorem 5.65 which asserts that every polynomial of positive degree having complex coefficients has a root in the complex field. Since there is no known really elementary proof of this theorem, we shall make use of results which are not proved in this book. In some sense this should be expected; in order to prove such a result, one must have information about the field of complex numbers which distinguishes it from other fields. It turns out, in fact, that it is properties of the real numbers which we shall require. The theorem which we quote without proof below is closely related to Theorem 3.38(ii) which was used earlier.

5.72 Definition. A ***real polynomial function*** of variables u and v is a function of the form

$$F(u,v) = a_{00} + a_{10}u + a_{01}v + \cdots + a_{ij}u^iv^j + \cdots + a_{mn}u^mv^n, \tag{5.73}$$

where the coefficients a_{ij} are real numbers.

The fact about real polynomial functions which we shall require is the following.

5.74 Theorem. *If $F(u,v)$ is a real polynomial function with the property $F(u,v) \geq 0$ for all real u and v, then there is some point (u_0,v_0) for which $F(u,v) \geq F(u_0,v_0)$ for all real u and v.*

This result is sometimes paraphrased by saying the nonnegative real polynomial function attains its minimum value at some point.

The next result gives the connection between complex polynomials and real polynomial functions which is needed for the proof of the main result of this section.

5.75 Theorem. *If $f(x)$ is a polynomial in $\mathbf{C}[x]$ and if we write a complex number z as $z = u + iv$, with u and v real, then there are real polynomial functions $a(u,v)$, $b(u,v)$ such that*

$$f(z) = f(u+iv) = a(u,v) + ib(u,v). \tag{5.76}$$

PROOF. The polynomial $f(x)$ is a sum of terms of the form cx^k for various complex numbers c and nonnegative integers k. Since the sum of two real polynomial functions of the form 5.73 is again a function of this form, it is enough to prove the statement for the case $f(x) = cx^k$. If $c = r + is$, then

$$f(z) = f(u+iv) = cz^k = (r+is)(u+iv)^k.$$

By expanding $(u+iv)^k$ with the aid of the binomial theorem, this is easily seen to be a real polynomial function plus i times a real polynomial function. When this is multiplied by the constant $r + is$, the result is again a sum of the form 5.76, and this is enough to prove the theorem.

We now restate Theorem 5.65 and complete its proof.

5.77 The Fundamental Theorem of Algebra. *If $f(x)$ is a polynomial of positive degree with complex coefficients, then there is a complex number z with $f(z) = 0$.*

PROOF. We assume the result is not true and will eventually reach a contradiction. Let $a(u,v)$ and $b(u,v)$ be the real polynomial functions which satisfy 5.76. Then the function $F(u,v)$ defined by

$$F(u,v) = |f(u + iv)|^2 = a(u,v)^2 + b(u,v)^2,$$

is a real polynomial function which satisfies $F(u,v) > 0$ for all real u, v. According to Theorem 5.74, there is some complex number $z_0 = u_0 + iv_0$ such that $F(u,v) \geq F(u_0, v_0)$ for all real u, v. Next we shall modify the original polynomial $f(x)$ and consider in its place a polynomial of the form $cf(px + q)$ for certain complex numbers c, p, q. These will be selected as follows: Take $c = 1/f(z_0)$; then the polynomial $|cf(u + iv)|^2$ has a minimum value 1 at $u_0 + iv_0$. We shall arrange for this minimum to occur at $z = 0$ by replacing the variable x with $px + z_0$, where p is a nonzero complex number to be selected later. If we set $h(x) = cf(px + z_0)$, with p a nonzero constant yet to be selected, then $h(0) = cf(z_0) = 1$ and $|h(z)| \geq 1$ for all complex numbers z. At this point we have

$$h(x) = cf(px + z_0) = 1 + b_k p^k x^k + \cdots + b_n p^n x^n$$

for some constants $b_k, \ldots, b_n$. We assume k is the smallest positive integer such that x^k has a nonzero coefficient. Then we select the number p to satisfy $p^k b_k = -1$; this is possible by Theorem 3.55. With these choices, the polynomial $h(x)$ can be written in the form

$$h(x) = 1 - x^k q(x),$$

where $q(x)$ is a polynomial in $\mathbf{C}[x]$ which satsifies $q(0) = 1$. We rewrite the above inequality as

$$1 \leq |h(z)|^2 = |1 - z^k q(z)|^2.$$

This may be interpreted geometrically as saying that the distance between $z^k q(z)$ and 1 is never less than 1, for any choice of the complex number z; equivalently, this says that the point $z^k q(z)$ always lies outside the circle C having radius 1 with center at $(1, 0)$. Now let $z = u$ take on real values. From the fact that $q(0) = 1$, we conclude that for sufficiently small values of the real variable u, the distance from $q(u)$ to 1 is small, say less than 0.5. (Here we are using the fact that polynomial functions are continuous.) This means that $q(u)$ lies inside the circle C. Since this is true for all sufficiently small real u, we may assume $0 < u < 1$. Since u is a positive real, the complex numbers $q(u)$ and $u^k q(u)$ have the same angle when written in trigonometric form in the sense of Definition 3.48. More precisely, the complex number $u^k q(u)$ lies on the straight line segment between the origin and the point $q(u)$. This segment lies entirely inside the circle C. This is a conflict with the assertion that $z^k q(z)$ lies outside C for all complex numbers z. In other words, the situation we have described cannot really occur. The assumption that the original polynomial had no root in **C** cannot be valid and the proof is complete.

Chapter 6

Groups

In all the algebraic systems studied so far we have always had two operations—namely, addition and multiplication. We now proceed to study an important class of systems in which there is only one operation. As soon as the definition of a *group* is given in the next section it will be apparent that we already have many examples of groups from previous chapters, although we have not used this terminology. The theory of groups is an important part of modern algebra, and many books have been written on the subject. In this chapter we shall present only a few of the most fundamental properties of groups and give a number of examples that may serve to suggest the wide range of applications of the theory.

6.1 DEFINITION AND SIMPLE PROPERTIES

Let "$\circ$" be a binary operation defined on a nonempty set G. We recall that this statement only means that if (a, b) is any ordered pair of elements of G, then $a \circ b$ is a uniquely determined element of G. As a matter of fact, it is customary to call this operation either "addition" or "multiplication," and to use the familiar notation that is associated with these words. However, we shall first give the definition of a group, using the symbol "$\circ$" for the operation.

6.1 Definition. A nonempty set G on which there is defined a binary operation* "$\circ$" is called a ***group*** (with respect to this operation) provided the following properties are satisfied:

---

* The assumption that "$\circ$" is a binary operation defined on G is sometimes expressed by saying that G is closed under this operation.

(i) If $a, b, c \in G$, then $(a \circ b) \circ c = a \circ (b \circ c)$ *(associative law).*
(ii) There exists an element e of G such that $e \circ a = a \circ e = a$ for every element a of G *(existence of an identity).*
(iii) If $a \in G$, there exists an element x of G such that $a \circ x = x \circ a = e$ *(existence of inverses).*

In Section 1.4 we called an element e, whose existence is asserted in (ii), an identity for the operation "$\circ$." However, we shall now call it an identity of the group. As suggested by the indicated name of the third property, as well as by previous use of the term, the element x whose existence is asserted in (iii) is called an *inverse* of the element a. Note that in a group *every* element has an inverse. As a matter of fact, it is quite easy to prove that in a group the identity is unique and also that every element has a unique inverse.

In order to give an example of a group, it is necessary to specify the elements of the set and to define an operation on this set in such a way that the three properties stated above are satisfied. We now give several examples of groups.

Example 1: The set **Z** of all integers, with the operation "$\circ$" taken as the usual operation (+) of addition. The first property then merely states that

$$(a + b) + c = a + (b + c),$$

and this is just the associative law of addition for the integers. In this case, the identity of the group is the zero integer since $0 + a = a + 0 = a$ for $a \in Z$. The inverse of the element a is the element $-a$ since $a + (-a) = -a + a = 0$, and in the present notation this is just what is required in the statement of the third property. We have therefore verified all three properties, and hence we have a group. This group may be called the *additive group of the integers.*

This example can easily be generalized as follows. Let S be the set of all elements of any *ring*, and let the operation "$\circ$" be taken as the operation of addition already defined in the ring. Then the three properties of a group are precisely the properties P_2, P_3, and P_4 of Section 2.2 required of addition in the ring. Hence, S must be a group relative to the operation of addition. We shall refer to this group as the *additive of the ring S.*

Example 2: The set T of all nonzero rational numbers, with the operation "$\circ$" taken as the familiar operation of multiplication of rational numbers. Since the product of two nonzero rational numbers is also a nonzero rational number, the set T is closed under multiplication; that is, multiplication is an operation defined on T. If $a, b, c \in T$, then $(ab)c = a(bc)$ by the associative law of multiplication for rational numbers, and this is just 6.1(i) in this case. Moreover, the identity is the rational number 1, and the inverse of an element a of T is the rational number a^{-1}. Hence, T is a group with respect to the operation of multiplication. We may emphasize that in a group *every* element must have an inverse, and this explains why the set of *all* rational numbers would not be a group with respect to multiplication.

This example can also be generalized as follows. If F is an arbitrary *field*, the set of all nonzero elements of F is a group with respect to the operation of multiplication in the field. This group we shall call the *multiplicative group of the field F*. We emphasize again that there is exactly one element of F, the zero, which is not an element of the multiplicative group of F.

Example 3: The set $\mathbf{R}^+$ of all *positive* real numbers with multiplication as the operation. Properties (ii) and (iii) are satisfied because $1 \in \mathbf{R}^+$ and if $a \in \mathbf{R}^+$, then also $a^{-1} \in \mathbf{R}^+$.

In like manner, $\mathbf{Q}^+$ is a group with respect to the operation of multiplication. More generally, if F is any *ordered* field, then F^+ is a group with respect to the operation of multiplication.

Example 4: The set $\{1, -1, i, -i\}$ consisting of these four complex numbers, with the operation of multiplication of complex numbers. It is easy to verify that the set is closed under multiplication and, of course, the number 1 is the identity. Moreover, 1 and -1 are their own inverses, and i and $-i$ are inverses of each other. The associative law clearly holds since it holds for multiplication of complex numbers in general.

Example 5: Let H be the set $\{p, q, r\}$ with an operation, which we shall consider as multiplication, defined by the following table.

$(\cdot)$	p	q	r
p	p	q	r
q	q	r	p
r	r	p	q

Clearly, p is the identity of H. Moreover, p is its own inverse, and q and r are inverses of each other. The associative law is also satisfied, although it would be tedious to verify it from the table.

Example 6: Let G be the set $\{e, a, b, c\}$ with an operation of multiplication defined by the following table.

$(\cdot)$	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

The associative law holds, although we shall not verify it; and e is the identity. In this group, each element is its own inverse.

Example 7: Let $U(\mathbf{Z}_n)$ be the set of all elements of the ring $\mathbf{Z}_n$ which have multiplicative inverses; these are the units of $\mathbf{Z}_n$ and are the cosets $u + (n)$ with u an integer relatively prime to n. The product of two such cosets is again a coset of this type, so $U(\mathbf{Z}_n)$ is a group.

This example can be generalized as follows. Let R be any ring with unity and let $U(R)$ be the set of all units of R. Then $U(R)$ is a group with identity equal to the unity element of R.

Example 8: Let $GL(2, \mathbf{Z})$ denote the set of all two by two matrices with integer entries

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

where the integers a, b, c, d satisfy $ad - bc = \pm 1$.

It can be shown that $GL(2, \mathbf{Z})$ is the set of units in the ring $M_2(\mathbf{Z})$ of all two by two matrices over the integers. Thus this example is a special case of the previous remark.

This example may also be generalized. If R is any commutative ring with unity, then $GL(2, R)$ is the set of all two by two matrices of the form g above in which $a, b, c, d \in R$ and $ad - bc \in U(R)$.

The notation GL may seem somewhat peculiar; it stands for the "general linear" group. This terminology is motivated by concepts which arise in the study of linear algebra. In that subject, one is most frequently interested in the case of matrices over a field. If F is a field, then $GL(2, F)$ is the set of matrices of the form g above in which $a, b, c, d \in F$ and $ac - bd \neq 0$. We shall consider this example for various choices of the field F in later sections.

The groups described in the first seven examples above have an additional property not required by the definition of a group. They are examples of abelian groups according to the following definition.

6.2 Definition. If in a group G with operation "$\circ$," $a \circ b = b \circ a$ for all $a, b \in G$, G is said to be an ***abelian group*** (or a commutative group).

The term *abelian group* is most commonly used for this concept. The name is derived from Niels Henrik Abel (1802–1829), a famous Norwegian mathematician whose fundamental work furnished an inspiration for many later mathematicians.

Most of the above examples are examples of abelian groups. In the following section we shall introduce some very important nonabelian groups different from those already introduced.

As in the examples, we shall always call the operation in a group either addition or multiplication, and shall use the usual notation associated with these names. We shall never use addition as the operation in a nonabelian group. That is, whenever addition is used as the operation, we shall always assume, whether or not it is explicitly mentioned, that the group is abelian. It follows that in such a group *all* the properties of addition in a ring are satisfied. The identity will be denoted by 0 and called "zero"; the inverse of an element a will be denoted by $-a$; we shall write $b - a$ for $b + (-a)$, and so on.

When the operation in a group is called multiplication, the group may be either abelian or nonabelian. Accordingly, when we come to prove a property of arbitrary groups, we shall think of the operation as multiplication, and use the implied notation. In particular, the inverse of an element a will then be denoted by a^{-1}. We shall usually let e be the identity of the group, and reserve the symbol "1" for the smallest positive integer.

Now let G be an arbitrary group with operation multiplication. The following properties can be easily proved using only trivial modifications of proofs that we have already met in our study of rings and fields. We may emphasize that here multiplication need not be commutative. The first two of these properties have already been stated in the preceding section.

6.3 Theorem. *The following hold in every group G:*

(*i*) *The identity of G is unique.*
(*ii*) *If $a \in G$, a has a unique inverse a^{-1}.*
(*iii*) *If $a, b, c \in G$ such that $ab = ac$, then $b = c$.*
(*iv*) *If $a, b, c \in G$ such that $ba = ca$, then $b = c$.*
(*v*) *If $a, b \in G$, there exists a unique element x of G such that $ax = b$, and a unique element y of G such that $ya = b$. In fact, $x = a^{-1}b$ and $y = ba^{-1}$.*
(*vi*) *The inverse of a product is the product of the inverses in the reverse order, that is, if $a, b \in G$, then $(ab)^{-1} = b^{-1}a^{-1}$.*

Properties (iii) and (iv) are naturally called the ***cancellation laws.*** The proofs of the various parts of 6.3 will be left as an exercise.

Just as in the case of multiplication in a ring, the generalized associative law holds, and we can write products without use of parentheses to indicate association.

If $a \in G$, we define $a^0 = e$, where e is the identity of the group. Then, just as though a were a nonzero element of a field, we can define a^n for *every* integer n. Moreover, for all choices of integers m and n, the following laws of exponents hold:

$$a^m \cdot a^n = a^{m+n},$$
$$(a^m)^n = a^{mn}.$$

For an *abelian* group, we also have $(ab)^n = a^n \cdot b^n$, but this is not true in general.

In those cases in which we use addition as the operation, we make use of multiples in place of powers; that is, na takes the place of a^n. Such a group is always assumed to be abelian and we have the following analogues of the above laws of exponents:

$$ma + na = (m + n)a,$$
$$n(ma) = (nm)a,$$
$$n(a + b) = na + nb.$$

These are properties that are already familiar as properties of addition in any ring.

Subgroups

A set H of elements of a group G is naturally called a *subgroup* of G if H is itself a group with respect to the operation already defined on G. If e is the identity of G, then G certainly has the two so-called *trivial* subgroups $\{e\}$ and G. Any other subgroup is called a *proper* subgroup.

The following theorem is often useful in determining subgroups of a given group.

6.4 Theorem. (*a*) *A nonempty subset K of a group G is a subgroup of G if and only if the following two conditions are satisfied:*

(*i*) *If $a, b \in K$, then $ab \in K$.*
(*ii*) *If $a \in K$, then $a^{-1} \in K$.*
(*b*) *If K has a finite number of elements, condition* (*ii*) *is implied by condition* (*i*).

PROOF. Condition (i) implies that K is closed under multiplication. If $a \in K$, then $a^{-1} \in K$ by (ii) and so by (i) we have $aa^{-1} = e$ is also in K. Thus K contains the identity element. The multiplication in all of G is associative, so in particular the multiplication in K is associative. All conditions in the definition of a group are met, so K is a subgroup of G when (i) and (ii) hold.

Now suppose that K is a finite set and (i) holds. We must show that for each $a \in K$, a^{-1} is an element of K. Consider the powers a, a^2, $a^3, \ldots$ of a; by (i) these are all in K. Since K is finite, there must be two integers $r > s > 0$ such that $a^r = a^s$. From this we conclude $a^{-1} = a^{r-s-1}$.

This element belongs to K so (ii) holds.

In view of this theorem, we see therefore that a nonempty set of elements of a group G having a finite number of elements is a subgroup of G if and only if the set is closed under the operation on G. However, for infinite groups condition (ii) is not a consequence of condition (i). As an example, the set of all positive integers is a nonempty subset of the additive group of the integers which satisfies condition (i) but not condition (ii).

We conclude this section by introducing one additional concept which is analogous to the concept of direct sum of rings. Suppose that G and H are groups and let us first assume that they are abelian groups and that the operation in each of these groups is written as addition. On the Cartesian product set $G \times H$ we define an operation of addition as follows:

$$(g_1, h_1) + (g_2, h_2) = (g_1 + g_2, h_1 + h_2), \qquad g_1, g_2 \in G;\ h_1, h_2 \in H.$$

It is easy to verify that we obtain in this way an abelian group. This group is called the *direct sum* of G and H and usually denoted by $G \oplus H$.

In a similar manner, if G and H are completely arbitrary groups with the operation in both groups written as multiplication, we may modify the above procedure in the obvious way by defining multiplication on the product set $G \times H$ as follows:

$$(g_1, h_1)(g_2, h_2) = (g_1 g_2, h_1 h_2), \qquad g_1, g_2 \in G;\ h_1, h_2 \in H.$$

The group obtained in this way is called the *direct product* of the groups G and H, and it is customary to denote it by the same notation $G \times H$ as used for the Cartesian product of the *sets* G and H.

We have defined the direct sum (or the direct product) of *two* groups. We leave it to the reader to give a corresponding definition of the direct sum (or the direct product) of any finite number of groups.

EXERCISES

1. Which of the following are groups with respect to the indicated operation?
   (a) The set $\{1, 3, 7, 9\}$ of elements of $\mathbf{Z}_{10}$, with operation multiplication.
   (b) The set $\{0, 2, 4, 6, 8\}$ of elements of $\mathbf{Z}_{10}$, with operation addition.
   (c) The set $\{1, 3, 9\}$ of elements of $\mathbf{Z}_{10}$, with operation multiplication.
   (d) The set of all rational numbers x such that $0 < x \leq 1$, with operation multiplication.
   (e) The set of all positive irrational real numbers with operation multiplication.
   (f) The set of all integers with operation "$\circ$" defined as follows: $a \circ b = a + b + 1$.
   (g) The set of all integers with operation "$\circ$" defined as follows: $a \circ b = a - b$.
   (h) The set of all rational numbers, other than 1, with operation "$\circ$" defined as follows: $a \circ b = a + b - ab$.
2. If H_1 and H_2 are subgroups of a group G, prove that $H_1 \cap H_2$ is a subgroup of G. Generalize this result by proving that the intersection of any number of subgroups of G is a subgroup of G.
3. Find all subgroups of each of the following groups:
   (a) The additive group of the ring $\mathbf{Z}_{12}$.
   (b) The additive group of the ring $\mathbf{Z}_5$.

(c) The multiplicative group of the field $\mathbf{Z}_7$.
(d) The multiplicative group of the field $\mathbf{Z}_{11}$.

4. Prove that $(ab)^2 = a^2b^2$ for all choices of a and b as elements of a group G if and only if G is abelian.

5. Let a be a fixed element of a group G. Prove that the set $\{x \mid x \in G, ax = xa\}$ is a subgroup of G.

6. Let G be the set of all ordered pairs (a, b) of real numbers with $a \neq 0$, and on this set let us define an operation of multiplication as follows:

$$(a, b)(c, d) = (ac, bc + d).$$

Verify that G is a nonabelian group.

7. Let $\mathbf{Z}_2$ be the field with two elements and $GL(2, \mathbf{Z}_2)$ the group of two by two matrices over $\mathbf{Z}_2$ defined in Example 8. Make a list of the six elements of this group. Show that the subset consisting of all elements of the form

$$\begin{bmatrix} a & 0 \\ c & d \end{bmatrix}$$

with $ad \neq 0$ forms a subgroup. Find at least two other subgroups of $GL(2, \mathbf{Z}_2)$ other than the trivial subgroups which consist of the identity alone or the entire group.

6.2 MAPPINGS AND PERMUTATION GROUPS

Suppose that A, B, and C are sets and that we have given mappings $\alpha: A \to B$ and $\beta: B \to C$. It is then easy to define in a natural way a mapping of A into C. If $a \in A$, we first take the image $\alpha(a)$ of a under the mapping α. Now $\alpha(a) \in B$, so we take the image of this element under the mapping β to get $\beta(\alpha(a))$ in C. Thus $a \to \beta(\alpha(a))$ defines a mapping of A into C determined by the two mappings α and β. We denote this mapping by $\beta\alpha$ and call it the product. More formally, the mapping $\beta\alpha: A \to C$ is defined by

6.5
$$\beta\alpha(a) = \beta(\alpha(a)), \qquad a \in A.$$

We may point out that, according to the definition just given, $\beta\alpha$ means "first perform α and then perform β." It is here that it makes an essential difference in notation whether we denote the image of a under α by $\alpha(a)$ or by the notation $a\alpha$. Had we adopted the latter notation, the mapping denoted by $\beta\alpha$ would map an element a of A into $a(\beta\alpha)$ of C; it would be most natural to interpret this as $(a\beta)\alpha$, which would mean "first perform β and then perform α." Both notations are used and in reading other books, the student must be prepared to find either one.

As a simple illustration of Definition 6.5, let $A = \{1, 2, 3, 4\}$, $B = \{x, y, z\}$, $C = \{r, s\}$, and let $\alpha: A \to B$ and $\beta: B \to C$ be defined, respectively, as follows:

$$\alpha(1) = y, \quad \alpha(2) = x, \quad \alpha(3) = x, \quad \alpha(4) = z;$$
$$\beta(x) = s, \quad \beta(y) = r, \quad \beta(z) = s.$$

Then the mapping $\beta\alpha: A \to C$ is obtained by the following calculations:

$$\begin{aligned}\beta\alpha(1) &= \beta(\alpha(1)) = \beta(y) = r,\\ \beta\alpha(2) &= \beta(\alpha(2)) = \beta(x) = s,\\ \beta\alpha(3) &= \beta(\alpha(3)) = \beta(x) = s,\\ \beta\alpha(4) &= \beta(\alpha(4)) = \beta(z) = s.\end{aligned}$$

It should be clear that in defining the product of two mappings a certain condition on the sets involved is necessary. Thus, if α is a mapping of A into B, then $\beta\alpha$ is defined only if β is a mapping of B into some set.

We now take one more step as follows. Let A, B, C, D be sets and suppose that we have mappings $\alpha: A \to B$, $\beta: B \to C$, $\gamma: C \to D$. Then $\beta\alpha$ is a mapping of A into C and $\gamma(\beta\alpha)$ is a mapping of A into D. In like manner, $(\gamma\beta)\alpha$ is a mapping of A into D. An important fact is that these two maps are equal, that is

6.6 $$\gamma(\beta\alpha) = (\gamma\beta)\alpha.$$

Let us prove this statement. One shows that two maps defined on A are the same by showing the two maps have the same value at each point of A. Thus it is required that we prove

6.7 $$[\gamma(\beta\alpha)](a) = [(\gamma\beta)\alpha](a)$$

for each element a of A.

First, we observe that by the definition of the product of the mappings γ and $\beta\alpha$, we have

$$[\gamma(\beta\alpha)](a) = \gamma[\beta\alpha(a)].$$

Then, by the definition of the product $\beta\alpha$ we have

$$\gamma[\beta\alpha(a)] = \gamma[\beta(\alpha(a))].$$

Thus the left side of 6.7 equals $\gamma[\beta(\alpha(a))]$. In like manner we obtain

$$[(\gamma\beta)\alpha](a) = (\gamma\beta)(\alpha(a)) = \gamma[\beta(\alpha(a))].$$

Since both sides of 6.7 are equal to $\gamma[\beta(\alpha(a))]$, we have proved 6.7 and also 6.6. Of course this merely states that *multiplication of maps is always associative.*

In connection with our study of groups we are interested in the special case of mappings of a set A *onto* the *same set* A. Moreover, it is the one-one mappings of A onto A that we wish to study. The following terminology is convenient.

6.8 Definition. A one-one mapping of a set A onto itself is called a ***permutation*** of the set A.

The next theorem will show why we have paused to study mappings in a discussion of groups.

6.9 Theorem. *The set S of all permutations of a set A is a group with respect to the operation of multiplication of mappings defined in 6.5.*

PROOF. First, let us give a formal proof of the essentially obvious fact that if α and β are permutations of A, then $\alpha\beta$ is also a permutation of A. Let x be any element of A. Since α and β are mappings of A onto A, there exists y in A with $\alpha(y) = x$; there exists z in A with $\beta(z) = y$. Then

$$\alpha\beta(z) = \alpha(\beta(z)) = \alpha(y) = x.$$

This shows $\alpha\beta$ maps A *onto* A. Next we show that $\alpha\beta$ is a one-one mapping, and therefore a permutation of A. It is necessary to verify that if a and b are elements of A such that $\alpha\beta(a) = \alpha\beta(b)$, then $a = b$. The equation $\alpha\beta(a) = \alpha\beta(b)$ can be rewritten as $\alpha(\beta(a)) = \alpha(\beta(b))$. Since α is one-one, it follows that $\beta(a) = \beta(b)$. Now since β is one-one, $a = b$, as required. This calculation shows that the set S is closed with respect to the operation of multiplication of permutations.

We have already proved that multiplication of mappings is associative, and hence the first requirement of a group is satisfied.

The identity mapping on the set A, as defined in Section 1.2, that is, the mapping $\epsilon: A \to A$ defined by

$$\epsilon(a) = a, \qquad a \in A,$$

is clearly an element of S and it is trivial to verify that $\alpha\epsilon = \epsilon\alpha = \alpha$ for every $\alpha \in S$. This shows that ϵ is the identity of our group. There only remains to prove that each element of S has an inverse.

If $\alpha \in S$, since α is a one-one mapping of A onto A, we have defined in Section 1.2 a one-one mapping α^{-1} of A onto A by the equation

$$\alpha^{-1}(\alpha(a)) = a, \qquad a \in A. \tag{6.10}$$

Thus $\alpha^{-1} \in S$ and, by our definition of the product of mappings, we have $\alpha^{-1}\alpha = \epsilon$. In order to show that α^{-1} is indeed the inverse of α, we must show that $\alpha\alpha^{-1} = \epsilon$. To verify this, let a be any element of A; then

$$\alpha\alpha^{-1}(\alpha(a)) = \alpha(\alpha^{-1}(\alpha(a))) = \alpha(\alpha^{-1}\alpha(a)) = \alpha(a).$$

Since every element b of A can be written as $b = \alpha(a)$ for some a, we have shown that

$$\alpha\alpha^{-1}(b) = b$$

for every b in A. Hence $\alpha\alpha^{-1} = \alpha^{-1}\alpha = \epsilon$ and α^{-1} defined by 6.10 is indeed the inverse of α.

We have now established all the properties required by the definition of a group, and the theorem is therefore established. This group S is naturally called the *group of all permutations of the set A*.

So far, the set A has been a completely arbitrary set. However, we are now primarily interested in the case in which A is restricted to have a finite number of elements. Accordingly, we make the following definition.

6.11 Definition. Let n be a positive integer. The group of all permutations of a set with n elements is called the ***symmetric group*** on n symbols, and may be designated by S_n.

Let us now consider an example in which $A = \{1, 2, 3\}$, a set with three elements. Then the symmetric group S_3, consisting of all permutations of A, contains six elements; we give examples of two of the elements α, β defined as follows:

6.12
$$\begin{aligned} &\alpha(1) = 2, \quad \alpha(2) = 1, \quad \alpha(3) = 3, \\ &\beta(1) = 2, \quad \beta(2) = 3, \quad \beta(3) = 1. \end{aligned}$$

The product of these two mappings may be computed by using the definition of the product of mappings. We have

$$\begin{aligned} \alpha\beta(1) &= \alpha(\beta(1)) = \alpha(2) = 1, \\ \alpha\beta(2) &= \alpha(\beta(2)) = \alpha(3) = 3, \\ \alpha\beta(3) &= \alpha(\beta(3)) = \alpha(1) = 2. \end{aligned}$$

This basic process of using the definition to compute products becomes very tedious when the number of elements in the set A is fairly large. Even simply writing down all the elements in the format 6.12 is rather time-consuming. An alternate notation may be used to describe the elements in S_n which makes the computation of products somewhat easier. We illustrate this notation with the elements of S_3. The mapping α defined above may be written as

6.13 $$\alpha = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$$

to express the fact that under the mapping α the image of 1 is 2, the image of 2 is 1, and the image of 3 is 3. According to this notation, we write the elements of the set A (in any order) in the top row, and under each element of A we write its image under the mapping α. In like manner the mapping β is written as

$$\beta = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.$$

Using this notation, one may compute the product efficiently. We illustrate the process. In order to compute $\alpha\beta$, first write the two permutations in the proper order using the double row notation and a partially blank permutation which will be the product:

$$\alpha\beta = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ a & b & c \end{pmatrix}.$$

The spaces now filled by a, b, c are left blank; our task is to find the correct elements of A to be placed there. To find the image of 1, that is a, we first look at the image of 1 in β and find it to be 2, and then find the image of 2 under α. The eye may be aided by following * → **, ** → *** in the product

$$\alpha\beta = \begin{pmatrix} 1 & 2^{**} & 3 \\ 2 & 1^{***} & 3 \end{pmatrix}\begin{pmatrix} 1^{*} & 2 & 3 \\ 2^{**} & 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & b & c \end{pmatrix}.$$

In this way one may quickly find $b = 3$, $c = 2$.

Using this notation makes it fairly easy to write down the six elements of S_3:

$$\epsilon = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \beta = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad \beta^2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix},$$

$$\alpha = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad \alpha\beta = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad \alpha\beta^2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}.$$

It may be noted that $\beta^3 = \alpha^2 = \epsilon$. We have used the double row notation to express each of the six elements and also have written each of these elements as a product involving α and β. Suppose we compute one more product:

$$(\alpha\beta^2)\alpha = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = \beta;$$

thus we have $(\alpha\beta^2)\alpha = \beta$. By similar computations, the entire multiplication table for this group may be determined with the elements expressed as the products of α and β listed above.

6.14

	ϵ	β	β^2	α	$\alpha\beta$	$\alpha\beta^2$
ϵ	ϵ	β	β^2	α	$\alpha\beta$	$\alpha\beta^2$
β	β	β^2	ϵ	$\alpha\beta^2$	α	$\alpha\beta$
β^2	β^2	ϵ	β	$\alpha\beta$	$\alpha\beta^2$	α
α	α	$\alpha\beta$	$\alpha\beta^2$	ϵ	β	β^2
$\alpha\beta$	$\alpha\beta$	$\alpha\beta^2$	α	β^2	ϵ	β
$\alpha\beta^2$	$\alpha\beta^2$	α	$\alpha\beta$	β	β^2	ϵ

It will be helpful to understand that even though the entry in the table that shows the product of $\alpha\beta^2$ and α ought to be written as $(\alpha\beta^2)\alpha$, the actual entry is β because that is the way the particular permutation is designated in the listing of all permutations given above. It would be good practice for the reader to verify the correctness of this table.

We can use a notation similar to 6.13 to denote a permutation of any finite set. In general, if $\{i_1, i_2, \ldots, i_n\}$ is an arrangement of the integers $1, 2, \ldots, n$, then by

$$\begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ i_1 & i_2 & i_3 & \cdots & i_n \end{pmatrix}$$

we mean the permutation α of the set $A = \{1, 2, \ldots, n\}$ such that $\alpha(1) = i_1, \alpha(2) = i_2, \ldots, \alpha(n) = i_n$. We shall use this notation whenever it seems convenient to do so.

We found above that S_3 has six elements. Let us now determine the number of elements in the symmetric group S_n, that is, the number of permutations of a set $A = \{1, 2, \ldots, n\}$ with n elements. Clearly, the image of 1 may be any element of A, and hence there are n choices for the image of 1. After an image of 1 is selected, there are then $n - 1$ choices for the image of 2, and so on. It follows that there are $n(n-1)\cdot(n-2)\cdots 2\cdot 1$ different permutations of A. This number is usually denoted by $n!$ and called "n factorial." We have therefore shown that S_n has $n!$ elements.

Any group whose elements are permutations is naturally called a *permutation group* or a *group of permutations*. Any subgroup of a symmetric group S_n is certainly a permutation group. For example, from the table 6.14 and Theorem 6.4 it follows that $\{\epsilon, \beta, \beta^2\}$ is a subgroup of S_3, and this is therefore an example of a permutation group which is not a symmetric group since it is not the group of *all* permutations of any set.

We conclude this section with a brief indication of how one can construct some interesting permutation groups by use of properties of symmetry of certain geometric figures. As an example, let us consider a square and study all rigid motions of the square into itself. That is, if the square is thought of as being made of some rigid material, such as cardboard, we consider motions such that the figure will look the same after the motion as before. In this, as well as in all other examples we shall consider, the rigid motions will consist of rotations either in the plane or in space. Each rigid motion of the square can be used in an almost obvious way to define a permutation of the vertices of the square. Let us designate the vertices of the square by 1, 2, 3, and 4. Moreover, let E, F, G, and H be the midpoints of the sides, as indicated in Figure 12; and let O be the center of the square. A rotation, in the plane of the square, through an angle of 90° about point O would place the vertices in the position shown in Figure 13. We may interpret the result of this rotation as mapping 1 into 2, 2 into 3, 3 into 4, and 4 into 1, that is, as effecting the permutation

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}$$

of the set $\{1, 2, 3, 4\}$ whose elements denote the vertices. A similar rotation through an angle of 180° or 270° leads to the respective permutations

$$\alpha^2 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} \quad \text{or} \quad \alpha^3 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix}.$$

Clearly, $\alpha^4 = \epsilon$, the identity permutation. We also have other rigid motions consisting of rotations in space about a line of symmetry of the square. Let β be the permutation which arises from a rotation through an angle of 180° about the line EF and γ the permutation which arises from a similar rotation about GH. Then we see that

$$\beta = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \quad \text{and} \quad \gamma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}.$$

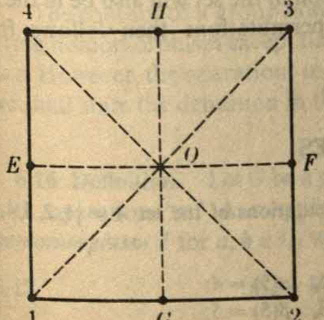


Figure 12

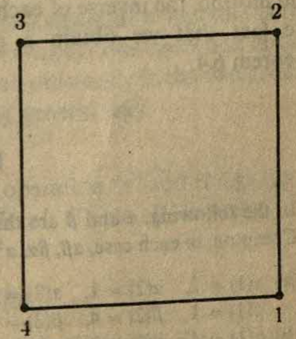


Figure 13

There remain two other permutations arising from rotations through 180° about the diagonals of the square. These are

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} \quad \text{and} \quad \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix}.$$

The set $\{\epsilon, \alpha, \alpha^2, \alpha^3, \beta, \gamma, \tau, \sigma\}$ of permutations obtained in this way is closed under multiplication, as is easily verified by the multiplication table given below.

6.15

	ϵ	α	α^2	α^3	β	γ	τ	σ
ϵ	ϵ	α	α^2	α^3	β	γ	τ	σ
α	α	α^2	α^3	ϵ	τ	σ	γ	β
α^2	α^2	α^3	ϵ	α	γ	β	σ	τ
α^3	α^3	ϵ	α	α^2	σ	τ	β	γ
β	β	σ	γ	τ	ϵ	α^2	α	α^3
γ	γ	τ	β	σ	α^2	ϵ	α^3	α
τ	τ	β	σ	γ	α^3	α	ϵ	α^2
σ	σ	γ	τ	β	α	α^3	α^2	ϵ

Moreover, it is evident that each permutation of this set has an inverse in this set, and we therefore have a group of permutations. This particular group with eight elements is called the *octic group*. Since we obtained this group by a consideration of the rigid motions of a square, we may also say that it is the group of rigid motions of a square.

In a similar way we may construct the group of rigid motions of other geometric figures. We observe that by this process we must actually obtain a *group* of permutations. In the first place, a rigid motion followed by another rigid motion is itself a rigid motion, and hence the set we obtain must be closed under multiplication. Since, also, each rigid motion can be reversed by another rigid motion, the inverse of each permutation in the set will also be in the set. The fact that we obtain a group of permutations then follows from Theorem 6.4.

EXERCISES

1. In the following, α and β are the given permutations of the set $A = \{1, 2, 3, 4, 5\}$. Compute, in each case, $\alpha\beta$, $\beta\alpha$, α^2, and β^2.

(a) $\alpha(1) = 2$, $\alpha(2) = 1$, $\alpha(3) = 3$, $\alpha(4) = 5$, $\alpha(5) = 4$;
$\beta(1) = 1$, $\beta(2) = 4$, $\beta(3) = 2$, $\beta(4) = 3$, $\beta(5) = 5$;
(b) $\alpha(1) = 4$, $\alpha(2) = 3$, $\alpha(3) = 5$, $\alpha(4) = 1$, $\alpha(5) = 2$;
$\beta(1) = 2$, $\beta(2) = 3$, $\beta(3) = 1$, $\beta(4) = 4$, $\beta(5) = 5$;

(c) $\alpha(1) = 2,\ \ \alpha(2) = 1,\ \ \alpha(3) = 4,\ \ \alpha(4) = 5,\ \ \alpha(5) = 3;$
$\beta(1) = 2,\ \ \beta(2) = 3,\ \ \beta(3) = 4,\ \ \beta(4) = 5,\ \ \beta(5) = 1.$

(d) $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 1 & 2 \end{pmatrix},\quad \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 5 & 4 \end{pmatrix}.$

(e) $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 2 & 5 & 4 \end{pmatrix},\quad \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 4 & 5 \end{pmatrix}.$

(f) $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 2 & 1 \end{pmatrix},\quad \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 2 & 1 & 3 \end{pmatrix}.$

2. Verify the entry in the table 6.15 giving each of the following products: $\alpha^2\sigma$, $\beta\gamma$, $\gamma\beta$, $\alpha^3\gamma$, $\tau\sigma$.

3. Find all six subgroups of the symmetric group S_3.

4. Find all ten subgroups of the octic group (6.15).

5. Show that the group of rigid motions of an equilateral triangle is the symmetric group S_3.

6. Find the group of rigid motions of a rectangle that is not a square. Make a multiplication table for this group and show that it is a subgroup of the octic group.

7. How many elements are there in the group of rigid motions of a regular pentagon? A regular hexagon?

8. Let $\alpha: A \to B$ be a given mapping. Prove each of the following:
   (i) If ϵ_A is the identity mapping on the set A, there exists a mapping $\beta: B \to A$ such that $\alpha\beta = \epsilon_A$ if and only if α is a one-one mapping.
   (ii) There exists a mapping $\gamma: B \to A$ such that $\gamma\alpha = \epsilon_B$ if and only if α is an onto mapping.
   (iii) If both the mappings β and γ exist, as defined in parts (i) and (ii), then $\beta = \gamma$.

6.3 HOMOMORPHISMS AND ISOMORPHISMS

The concept of a group homomorphism is essentially the same as that of a ring homomorphism except that we now have only one operation instead of two. However, the operations may be written differently in the two groups, so we shall state the definition in the following general way.

6.16 Definition. Let G be a group with operation "$\bullet$" and H a group with operation "$*$." A mapping $\theta: G \to H$ of G into H is called a ***homomorphism*** if for $a, b \in G$, we have

6.17
$$\theta(a \bullet b) = \theta(a) * \theta(b).$$

If there is a homomorphism of G *onto* H, we may say that H is a ***homomorphic image*** of G.

The special case in which the mapping is one-one and onto is of such importance that we introduce the following terminology which has already been used for rings.

6.18 Definition. A homomorphism which is a one-one and onto mapping is called an ***isomorphism.*** If there exists an isomorphism of G onto H, we say that G is isomorphic to H or that H is an isomorphic image of G.

Of course, just as with rings, if θ is an isomorphism of G onto H, then θ^{-1} is an isomorphism of H onto G, and we may say that G and H are isomorphic.

The operations "$\circ$" and "$*$" appearing in Definition 6.16 will be considered to be either addition or multiplication, but the new feature here is that one of them may be written as addition and the other as multiplication. In any case, we may for convenience indicate that condition 6.17 holds by saying that *the group operation is preserved under the mapping* θ.

Let us now illustrate these concepts by a few illustrative examples.

Example 1: Let G be the additive group of the ring $\mathbf{Z}_4$ and let H be the additive group of the ring $\mathbf{Z}_8$. For convenience let us designate the elements of G by 0, 1, 2, 3, and the elements of H by $0^*, 1^*, \ldots, 7^*$. Now let $\theta: G \to H$ be defined by

$$\theta(0) = 0^*, \quad \theta(1) = 2^*, \quad \theta(2) = 4^*, \quad \theta(3) = 6^*.$$

We assert that the group operations are preserved by this mapping θ. We shall not verify this but simply give an illustration in one case:

$$\theta(2 + 3) = \theta(5) = \theta(1) = 2^*,$$
$$\theta(2) + \theta(3) = 4^* + 6^* = 10^* = 2^*.$$

The reader may verify that θ is a one-one homomorphism of G into H; it is not onto H, however, since, for example, there is no element $a \in G$ with $\theta(a) = 1^*$.

Example 2: Let $\mathbf{R}^+$ be the group of all positive real numbers with operation multiplication (Example 3 of Section 6.1) and let L be the additive group of the field of all real numbers. If $\phi: \mathbf{R}^+ \to L$ is defined by

$$\phi(x) = \log_{10} x, \qquad x \in \mathbf{R}^+,$$

it is known that ϕ is a one-one mapping of $\mathbf{R}^+$ onto L and, moreover, one of the familiar laws of logarithms assures us that

$$\phi(xy) = \log_{10}(xy) = \log_{10} x + \log_{10} y = \phi(x) + \phi(y), \qquad x, y \in \mathbf{R}^+.$$

This shows that the group operation is preserved under the mapping ϕ, and ϕ is therefore an isomorphism of $\mathbf{R}^+$ onto L. It is this fact that is of central importance in the theory of logarithms. Of course, in place of 10 we could use as a base any fixed element of $\mathbf{R}^+$ other than 1.

Example 3: Let S_3 be the symmetric group on three symbols, with multiplication table given by 6.14, and let H be the group consisting of the set $\{1, -1, i, -i\}$ of complex numbers with operation multiplication. Let $\theta: S_3 \to H$ be defined as follows:

$$\theta(\epsilon) = \theta(\beta) = \theta(\beta^2) = 1, \quad \theta(\alpha) = \theta(\alpha\beta) = \theta(\alpha\beta^2) = -1.$$

It may be verified that the group operation is preserved under this mapping. As an example, we have

$$\theta(\alpha\beta^2 \cdot \alpha) = \theta(\beta) = 1,$$
$$\theta(\alpha\beta^2)\theta(\alpha) = (-1)(-1) = 1.$$

Thus θ is a homomorphism of S_3 into H. However, θ is clearly not an onto mapping, so we cannot say that H is a homomorphic image of S_3. However, if we let $H' = \{1, -1\}$, then H' is a subgroup of H and θ does define a mapping of S_3 onto H', and hence H' is a homomorphic image of S_3.

Example 4: Let G and H be arbitrary groups with operation multiplication. Then the mapping $\theta: G \times H \to G$ defined by $\theta(g, h) = g$ (that is, the projection of the set $G \times H$ onto G) is clearly a homomorphism of the direct product $G \times H$ onto G. Similarly, there exists a homomorphism of $G \times H$ onto H. Thus, both of the groups G and H are homomorphic images of the direct product $G \times H$.

We may point out that there always exists a (trivial) homomorphism of an arbitrary group G into an arbitrary group H in which every element of G maps into the identity of H.

Some fundamental properties of homomorphisms of groups are stated in the following theorem (cf. Theorem 2.33), in which we shall consider the operations of both groups to be multiplication.

6.19 Theorem. *Let $\theta: G \to H$ be a homomorphism of the group G into the group H. Then each of the following is true:*

(*i*) *If e is the identity of G, then $\theta(e)$ is the identity of H.*
(*ii*) *If $a \in G$, then $\theta(a^{-1}) = \theta(a)^{-1}$.*
(*iii*) *If G is abelian and θ is an onto mapping, then H is abelian.*

We leave the proof of this theorem as an exercise.

The reader may sometimes have to make a suitable modification in notation. For example, if the operation in G is multiplication and that in H is addition, property (ii) of the preceding theorem should be interpreted as saying that $\theta(a^{-1}) = -\theta(a)$, since $-\theta(a)$ is the inverse of $\theta(a)$ in H.

The concept introduced in the following definition plays an important role in the study of homomorphisms of groups.

6.20 Definition. Let $\theta: G \to H$ be a homomorphism of the group G into the group H. The set of all elements of G which map into the identity of H is called the ***kernel*** of the homomorphism θ, and may be denoted by $\ker \theta$.

In Example 3 of this section, $\ker \theta = \{\epsilon, \beta, \beta^2\}$. This is a subgroup of S_3. The next theorem shows that this will always be the case.

6.21 Theorem. *If $\theta: G \to H$ is a homomorphism, then $\ker \theta$ is a subgroup of G. Moreover, if e is the identity of G, $\ker \theta = \{e\}$ if and only if θ is a one-to-one mapping.*

PROOF. Let $K = \ker \theta$. By Theorem 6.19(i), the identity of H is $\theta(e)$. If $a, b \in K$, then $\theta(ab) = \theta(a)\theta(b) = \theta(e)\theta(e) = \theta(e)$, and thus K is closed with respect to multiplication. Moreover, if $a \in K$, by Theorem 6.19(ii) we have $\theta(a^{-1}) = \theta(a)^{-1} = \theta(e)^{-1} = \theta(e)$ and hence $a^{-1} \in K$. Theorem 6.4 then shows that K is a subgroup of G.

To prove the second statement of the theorem, suppose first that θ is one-to-one. Since $e \in \ker \theta$, and the mapping is one-to-one, this ensures that no other element of G has as image the identity $\theta(e)$ of H. It follows that $\ker \theta = \{e\}$. Conversely, let us suppose that $\ker \theta = \{e\}$. If $a, b \in G$ are elements with $\theta(a) = \theta(b)$, then

$$\theta(ab^{-1}) = \theta(a)\theta(b)^{-1} = \theta(a)\theta(a)^{-1} = \theta(e).$$

This shows that $ab^{-1} \in K$ and thus $ab^{-1} = e$ and $a = b$. This shows that θ is a one-to-one mapping. This completes the proof of the theorem.

The result just established is often useful in the following way. To show that a mapping of a group G into a group H is an isomorphism, we may first show that it is a homomorphism of G onto H and then verify that its kernel consists of the identity only.

We conclude this section by proving a theorem which is due to the famous English mathematician Arthur Cayley (1821–1895).

6.22 Theorem. *Every group G is isomorphic to a group of permutations.*

PROOF. In order to prove this result we need first of all to determine the *set* some of whose permutations we shall associate with the elements of the given group G. We make what is perhaps the most obvious choice, namely, the set of elements of G itself. Moreover, we shall let G denote both the group and the set of its elements as it suits our convenience. Actually, the desired permutations will be obtained by multiplication by the elements of the group. More precisely, let us first observe that if $a \in G$, then

$$\{ax \mid x \in G\} = G.$$

Therefore the mapping $\theta_a : G \to G$ defined by $\theta_a(x) = ax$ is a mapping of G onto G. It is easily seen to be one-one and so θ_a is a permutation of the set G which is associated with the element a of the group G.

Now let us set

$$H = \{\theta_a \mid a \in G\};$$

that is, H is the set of all permutations of the type introduced above associated with the elements of G. Now for every $a, b, x \in G$, we have

$$(\theta_a \cdot \theta_b)(x) = \theta_a(\theta_b(x)) = \theta_a(bx) = a(bx) = \theta_{ab}(x),$$

we see that

6.23 $$\theta_a \cdot \theta_b = \theta_{ab},$$

and H is therefore closed with respect to multiplication. Moreover, H has identity θ_e, where e is the identity of G; and 6.23 shows that θ_a has inverse equal to θ_b if $b = a^{-1}$. It follows that H is a subgroup of the group of all permutations of the set G.

We now assert that the mapping $\alpha : G \to H$ defined by $\alpha(a) = \theta_a$ is an isomorphism of G with H. It is clearly an onto mapping. Moreover, it is a homomorphism since, by 6.23,

$$\alpha(ab) = \theta_{ab} = \theta_a \cdot \theta_b = \alpha(a)\alpha(b).$$

There remains only to prove that α is one-to-one. Suppose $\alpha(a) = \alpha(b)$; that is, $\theta_a = \theta_b$. Then these two functions have the same value at each point of the set G; in particular, at the identity we find

$$a = \theta_a(e) = \theta_b(e) = b.$$

Thus α is indeed a one-one mapping, and the proof is complete.

In view of this theorem, in order to prove a result about all groups, it is sufficient to prove it for groups of permutations. Although the subject of group theory is not usually approached in this way, we shall see in the next chapter how the study of finite groups is facilitated by ingenious use of certain groups of permutations.

EXERCISES

1. Verify that there exists an isomorphism of the additive group of the ring $\mathbf{Z}_6$ onto the multiplicative group of the field $\mathbf{Z}_7$ such that the image of the element 1 of the first group is the element 3^* of the second group.

2. Find four different homomorphisms of the additive group $\mathbf{Z}_4$ into the additive group $\mathbf{Z}_8$.

3. Prove that a group G is abelian if and only if the mapping $\theta: G \to G$ defined by $\theta(a) = a^{-1}$, $a \in G$, is an isomorphism.

4. If G is the additive group of the ring $\mathbf{Z}_{15}$ and H is the additive group of the ring $\mathbf{Z}_5$, find a homomorphism of G onto H.

5. Show that two groups are necessarily isomorphic if each of them has exactly two elements. Show that the same conclusion holds if each has exactly three elements.

6. Find two groups, each with exactly four elements, which are not isomorphic. Do the same thing, with "four" replaced by "six."

7. If G is the multiplicative group of the field $\mathbf{Z}_5$, use the method of proof of Theorem 6.22 to find a group of permutations which is isomorphic to G.

8. If $\theta: G \to H$ and $\phi: H \to K$ are group homomorphisms, show that $\phi\theta$ is also a homomorphism.

9. The subgroup $\{\epsilon, \alpha^2, \beta, \gamma\}$ of the octic group (6.15) is sometimes called the *four-group*. Verify that the four-group is isomorphic to the additive group of the ring $\mathbf{Z}_2 \oplus \mathbf{Z}_2$, and also to the group of Example 6 of Section 6.1.

10. If $\theta: G \to H$ is a homomorphism of G into H and $a \in G$, prove that $\theta(a^k) = \theta(a)^k$ for *every* integer k.

11. Let G be the set of all two by two matrices of the form

$$g = \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}, \quad a, b \in \mathbf{R}, a \neq 0.$$

Show that G is a group with the operation of matrix multiplication. Furthermore, show that G contains a subgroup isomorphic to the multiplicative group of nonzero real numbers and also a subgroup isomorphic to the additive group of real numbers.

12. For each ordered pair (a, b) of real numbers with $a \neq 0$, let $\alpha_{a,b}: \mathbf{R} \to \mathbf{R}$ be the mapping of the field $\mathbf{R}$ into the field $\mathbf{R}$ defined by $\alpha_{a,b}(x) = ax + b$, $x \in \mathbf{R}$. Prove that the set H of all such mappings $\alpha_{a,b}$ is a group of permutations of $\mathbf{R}$, and that H is isomorphic to the group G mentioned in the preceding exercise.

13. Determine the set L of all elements $\alpha_{a,b}$ of the group H of the preceding exercise such that $\alpha_{a,b}\alpha_{1,2} = \alpha_{1,2}\alpha_{a,b}$. How do you know without detailed calculation that L is a subgroup of H? Verify that the group L is isomorphic to the additive group of the field $\mathbf{R}$.

14. Let G be a group and H merely a *set* on which a binary operation is defined. If there exists a mapping $\theta: G \to H$ of G *onto* H which preserves the operation, prove that H is a group with respect to the given operation and that θ is a homomorphism of the group G onto the group H. [This fact is usually expressed by saying that a homomorphic image of a group is a group.]

15. Let G be the multiplicative group of the field $\mathbf{Q}$, and let H be the set of all rational numbers other than 1. If $a, b \in H$, let us define $a \circ b = a + b - ab$. Prove that H is a group with respect to this operation. Prove also that the mapping $\theta: G \to H$ defined by $\theta(a) = 1 - a, a \in G$, is an isomorphism of G onto H.

6.4 CYCLIC GROUPS

If a is an element of an arbitrary group G, then since G is closed with respect to the operation (which we will consider to be multiplication), we see that $a^k \in G$ for every positive integer k. Moreover, a^0 is the identity e of G by definition, and a^{-k} is the inverse of a^k. It follows easily that the set $\{a^k \mid k \in \mathbf{Z}\}$ is a subgroup of G. We are particularly interested in the case in which this subgroup turns out to be all of G. Accordingly, let us make the following definition.

6.24 Definition. If the group G contains an element a such that $G = \{a^k \mid k \in \mathbf{Z}\}$, we say that G is a ***cyclic group*** and that G is ***generated by a*** or that a is a ***generator*** *of* G.

Since $a^i \cdot a^j = a^j \cdot a^i$ for $i, j \in \mathbf{Z}$, we see that a cyclic group is necessarily abelian.

Whether or not a group G is cyclic, if $a \in G$, the subgroup $\{a^k \mid k \in \mathbf{Z}\}$ of G is a cyclic group which we naturally call the *subgroup of G generated by a*.

Let us now give some examples of cyclic groups.

Example 1: The multiplicative group of the field $\mathbf{Z}_5$. Let us write the elements as 1, 2, 3, and 4; and remember that multiplication is to be carried out modulo 5. It is easily verified that $2^1 = 2, 2^2 = 4, 2^3 = 3, 2^4 = 1$; hence each element of the group is of the form 2^k for some integer k. It follows that the group must be cyclic with generator 2. The reader may show that 3 is also a generator of this group. The element 4 generates the cyclic subgroup $\{1, 4\}$.

Example 2: The additive group of the ring $\mathbf{Z}$ of integers. In a ring with addition as the operation, ka is the analogue of a^k used above. The integer 1 is a generator of this group since every element is of the form $k \cdot 1$ for some integer k.

Example 3: The additive group of the ring $\mathbf{Z}_n$ of integers modulo n. This group is generated by the element 1 of $\mathbf{Z}_n$. Of course, it may have other generators as well.

Example 4: The subset $\{\epsilon, \beta, \beta^2\}$ of the symmetric group S_3, whose multiplication table is given in 6.14 is a cyclic group. It is seen from that table that $\beta^3 = \epsilon$ and so the set is closed under multiplication. The group S_3 is not cyclic. This may be seen by direct calculation or, more simply, by noticing that S_3 is nonabelian, whereas a cyclic group is always abelian.

Example 5: Let α be an element of any symmetric group, S_n. The set $\{\alpha, \alpha^2, \ldots\}$ consisting of all powers of α is certainly closed under multiplication. Since S_n is a finite group, the set of powers of α is a subgroup of S_n. By definition this is a cyclic subgroup.

We are now ready to give another definition as follows:

6.25 Definition. (i) If a group G has n elements, where n is a positive integer, G is said to have ***finite order*** or, more precisely, to have ***order n***. If there exists no such positive integer, G is said to have ***infinite order***.

(ii) The ***order of an element a*** of a group G is the order of the cyclic subgroup of G generated by a.

In the language introduced here, we may say that the additive group of the integers has infinite order, the symmetric group S_n has order $n!$, the additive group of the ring $\mathbf{Z}_n$ has order n, and the multiplicative group of the field $\mathbf{Z}_p$ has order $p - 1$. All of these, except the first, are groups of finite order.

The next theorem gives an important characterization of the order of an element of a group. In fact, in order to compute the order of a given element it is usually simpler to apply this theorem than to use the definition.

6.26 Theorem. *An element a of a group G has order n if and only if n is the smallest positive integer such that* $a^n = e$, *where e is the identity of G. If no such integer exists, a has infinite order.*

As a first step in the proof, we shall prove the following lemma.

6.27 Lemma. *Let a be an element of the group G, and suppose that* $a^n = e$, *with n the smallest such positive integer. If* $k \in \mathbf{Z}$, *then* $a^k = e$ *if and only if* $k \equiv 0 \pmod{n}$. *More generally, if* $i, j \in \mathbf{Z}$, then $a^i = a^j$ *if and only if* $i \equiv j \pmod{n}$.

PROOF. By the Division Algorithm, we may write any integer k in the form $k = qn + r$, where q and r are integers and $0 \leq r < n$. Then, since

$a^n = e$, we have

$$a^k = a^{qn+r} = (a^n)^q \cdot a^r = e^q \cdot a^r = a^r.$$

If $a^k = e$, we see that $a^r = e$, and it follows that $r = 0$ since, otherwise, r would be a positive integer less than n and we have assumed that n is the smallest positive integer such that $a^n = e$. Hence, if $a^k = e$, we have $k = qn$ and $k \equiv 0 \pmod{n}$. Conversely, if $k = qn$, it is apparent that $a^k = (a^n)^q = e^q = e$. This establishes the first part of the lemma. The second part now follows easily. For if $a^i = a^j$, it follows that $a^{i-j} = e$, and by what we have just proved, this is true if and only if $i - j \equiv 0 \pmod{n}$ or $i \equiv j \pmod{n}$.

Let us return to the proof of the theorem, and suppose first that $a^n = e$, with n as the smallest such positive integer. We now assert that the elements

6.28 $$e, a, a^2, \ldots, a^{n-1}$$

are distinct and are all of the elements of the cyclic subgroup of G generated by a. Since no two of the integers $0, 1, 2, \ldots, n-1$ are congruent modulo n, the preceding lemma shows that the elements 6.28 are distinct. Moreover, since every integer is congruent modulo n to some one of the integers $0, 1, 2, \ldots, n-1$, it also follows that a^k is equal to one of the elements 6.28, for every integer k. Hence the cyclic subgroup of G generated by a has exactly the n distinct elements 6.28; that is, it has order n and therefore a has order n.

To prove the converse, let us now assume that a has order n. Then not all positive powers of a can be distinct; that is, we must have $a^i = a^j$ for different positive integers i and j. Suppose that $i > j$, and it then follows that $a^{i-j} = e$, with $i - j > 0$. Hence, there exists some positive power of a which is equal to e. Suppose that m is the smallest positive integer such that $a^m = e$. Now, by what we have proved above, a has order m. Since it was given that a has order n, we must have $m = n$. This completes the proof of the first sentence of the theorem.

If there exists no positive integer n such that $a^n = e$, it is easy to show that $a, a^2, a^3, \ldots$ must all be distinct. (Why?) Hence a must have infinite order, and the theorem is established.

Theorem 6.26 makes it easy to determine the order of an element of a given group. For example, let us find the order of the element 3 of the multiplicative group of the ring $\mathbf{Z}_{11}$. By computing the successive powers of 3, we find that $3^2 = 9$, $3^3 = 5$, $3^4 = 4$, $3^5 = 1$. Hence, the element 3 has order 5. As another example, let us find the order of the element 10 of the additive group of the ring $\mathbf{Z}_{18}$. By Theorem 6.26, with the proper change of notation, this order will be the least positive integer n such that $n \cdot 10 \equiv 0 \pmod{18}$. It follows easily that $n = 9$, and the element 10 therefore has order 9.

In view of Theorems 6.26 and 5.35, we see that if a ring R has a unity e and has nonzero characteristic, the characteristic of R is simply the order of e in the additive group of R.

It is not difficult to prove that two cyclic groups are isomorphic if and only if they have the same order. This fact will follow immediately from the following theorem.

6.29 Theorem. (*i*) *Every cyclic group of infinite order is isomorphic to the additive group of the ring* **Z** *of integers.*
(*ii*) *Every cyclic group of order n is isomorphic to the additive group of the ring* $\mathbf{Z}_n$ *of integers modulo n.*

PROOF. Let G be a cyclic group with generator a and let $\theta: \mathbf{Z} \to G$ be the mapping defined by

$$\theta(k) = a^k, \qquad k \in \mathbf{Z}.$$

if $i, j \in \mathbf{Z}$, then

$$\theta(i + j) = a^{i+j} = a^i a^j = \theta(i)\theta(j);$$

thus θ is a homomorphism of the additive group of **Z** to the group G. In fact, θ maps **Z** onto G because every element of G is a power of a. Now we consider two cases. If G has infinite order, then there is no positive integer n such that $a^n = e$, $e =$ the identity of G. This means there is no positive integer in the kernel of θ. Since $\ker \theta$ is a subgroup of **Z**, there is no negative integer in $\ker \theta$ either, and so $\ker \theta = \{0\}$ and θ is one-one. It follows that θ is an isomorphism of **Z** with G; that is, in the case that G has infinite order, G is isomorphic to the additive group of integers.

Now suppose G has finite order n. Then n is the smallest positive integer for which $a^n = e$. It follows that $a^k = e$ if and only if k is a multiple of n. This means that $\ker \theta = (n)$, the set of all multiples of n. The distinct elements $a, a^2, \ldots, a^n = e$, correspond in a one-to-one way with the cosets $1 + (n)$, $2 + (n), \ldots, n + (n) = 0 + (n)$. We define a mapping from the set of these cosets, which is the additive group of the ring $\mathbf{Z}_n$, by the rule $i + (n) \to a^i$; since this can be expressed as the mapping $s + (n) \to \theta(s)$, for any integer s, it is easily seen that this mapping is well-defined, and it is a one-one homomorphism of $\mathbf{Z}_n$ onto G; thus these two groups are isomorphic.

It may be useful to point out that an isomorphism from the additive group Z_n to a group G is completely determined once the image of the coset

$1 + (n)$ is specified. By this we mean that there is only one homomorphism which sends $1 + (n)$ to a specific element b. Once this has been specified, then for any positive integer k, we have

$$k + (n) = k[1 + (n)] \to b^k$$

and so the image of every element of $\mathbf{Z}_n$ is determined.

As an example, consider the multiplicative group of the field $\mathbf{Z}_5$. It is a group with four elements and the powers of the coset containing 2 are the elements $[2], [4], [2^3] = [3], [2^4] = [1]$. Since these are the four elements of this multiplicative group, the group is cyclic with $[2]$ as generator. The map of the additive group of $\mathbf{Z}_4$ onto this multiplicative group is given by

$$i + (4) \to [2^i].$$

Written out in full detail, we have

$$1 + (4) \to [2], \quad 2 + (4) \to [4], \quad 3 + (4) \to [3], \quad 0 + (4) \to [1].$$

There is another isomorphism between these groups which can be obtained by using another generator of the multiplicative group. The class $[3]$ is also a generator in this example, so the mapping defined by $i + (4) \to [3^i]$ provides a different isomorphism.

Our final theorem about cyclic groups is the following.

6.30 Theorem. *Every subgroup of a cyclic group G is itself a cyclic group.*

PROOF. Suppose that G is generated by a, and let H be a subgroup of G. Let m be the smallest positive integer such that $a^m \in H$. We shall show that H is a cyclic group generated by a^m. Since $H \subseteq G$, any element of H is of the form a^k for some integer k. By the Division Algorithm, we may write $k = qm + r$, where $0 \leq r < m$. Hence,

$$a^k = a^{qm+r} = (a^m)^q \cdot a^r,$$

and from this it follows that

$$a^r = (a^m)^{-q} \cdot a^k.$$

Since $a^m \in H$ and $a^k \in H$, this equation implies also that $a^r \in H$. In view of the choice of m as the smallest positive integer such that $a^m \in H$, and since $r < m$, we must have $r = 0$. We conclude that $k = qm$, and hence that every element a^k of H is of the form $(a^m)^q$ for some integer q. This shows that H is a cyclic group generated by a^m.

As an almost immediate consequence of the *proof* of the preceding theorem, we obtain the following result.

6.31 Corollary. *If a cyclic group G has finite order n and is generated by a, every subgroup H of G is generated by an element of the form a^m, where m is a divisor of n.*

PROOF. Since Theorem 6.26 shows that $a^n = e$, and $e \in H$, we apply the above argument with $k = n$ and obtain $n = qm$. Hence m is a divisor of n.

Consider the cyclic group of order 18 which has generator a. It has a subgroup consisting of all powers of the element a^{10}. However, since 10 does not divide 18, this subgroup must have a different generator of the form a^k, with k a divisor of 18. In fact, the proof of 6.30 indicates that k may be selected as the least positive integer with the property that $a^k = a^{10t}$, for some integer t: that is, a^k is equal to some element of the subgroup consisting of all powers of a^{10}. A little calculation will show that a is not in this subgroup but a^2 is in the subgroup; thus the subgroup consisting of all powers of a^{10} is generated not only by a^{10} but also by a^2. It is no coincidence that 2 happens to be the g.c.d. of 10 and 18. We leave as an exercise for the reader the proof of the following general statement.

6.32 Corollary. *If G is a cyclic group of order n with generator a, then the subgroup of G generated by a^r is also generated by a^d where $d = (r, n)$. In particular, the elements which generate G are those a^k with $(k, n) = 1$.*

Of course, by a simple change in notation, these results apply equally well to the case in which the operation is addition. As an illustration of the preceding corollary, let us find all subgroups of the additive group of the ring $\mathbf{Z}_{14}$. This is a cyclic group of order 14 generated by the element 1; hence the only subgroups are the cyclic subgroups generated by 1, 2, 7, and 14. The subgroup generated by 14 consists only of the identity 0. The subgroup generated by 2 has order 7 and the subgroup generated by 7 has order 2. The subgroup of $\mathbf{Z}_{14}$ generated by 8 is the same as the subgroup generated by $(8, 14) = 2$—namely, the subgroup of order 7.

EXERCISES

1. Find the order of each element of the octic group (7.15).
2. Find an element of the symmetric group S_4 of order 4. Similarly, find an element of S_5 of order 5; of S_n of order n.

3. It will be proved later that for every prime p, the multiplicative group of the field $\mathbf{Z}_p$ is cyclic. Verify this statement now for the primes $p = 7, 11$, and 13.
4. Find all subgroups of the additive group of the ring $\mathbf{Z}_{20}$.
5. Prove: If G is a cyclic group of order n and k is a positive divisor of n, there exists a subgroup of G of order k.
6. Prove the two statements in Corollary 6.32.
7. Find all generators of the additive group $\mathbf{Z}_{30}$.
8. Prove that if $\theta: G \to H$ is a homomorphism of a cyclic group G with generator a onto a group H, then H is a cyclic group with generator $\theta(a)$ and that if G has finite order, the order of $\theta(a)$ is a divisor of the order of a.
9. Let G and H be cyclic groups of the same order, and let g be an arbitrary generator of G and h an arbitrary generator of H. Show that there exists an isomorphism θ of G onto H such that $\theta(g) = h$.
10. Determine all isomorphisms of the multiplicative group of the field $\mathbf{Z}_{11}$ onto the additive group of the ring $\mathbf{Z}_{10}$.
11. Determine all subgroups of the additive group of the ring $\mathbf{Z}$.
12. Let G be a cyclic group of order n and let H be a cyclic group of order m. Prove that, if m is a divisor of n, then there is a homomorphism of G onto H.
13. Let G be a cyclic group of order n and H a cyclic group of order m. Prove that a subgroup K of G is isomorphic to some subgroup of H if and only if the order of K divides the g.c.d. of n and m.

6.5 COSETS AND LAGRANGE'S THEOREM

Let G be an arbitrary group and H a subgroup of G. If $a \in G$, we shall designate by aH the *set* of all elements of G of the form ah, where $h \in H$. That is, $aH = \{ah \mid h \in H\}$.

6.33 Definition. If H is a subgroup of the group G and $a \in G$, we call aH a ***coset*** of H (in G).*

Since $eH = H$, we see that H is itself a coset. Moreover, since $e \in H$, it is clear that $a \in aH$.

---

* More precisely, we have here defined a *left* coset, and one can similarly define a right coset Ha. However, in accordance with the definition just given we shall in this section use the word *coset* to mean *left coset*.

The following lemma will be very useful in studying cosets.

6.34 Lemma. *If H is a subgroup of the group G and $a, b \in G$, then each of the following is true:*

(*i*) *If $aH \cap bH \neq \varnothing$, then $aH = bH$.*
(*ii*) *$aH = bH$ if and only if $a \in bH$.*

PROOF OF (i). Suppose that aH and bH have at least one element in common. Thus, there exist $h_1, h_2 \in H$ such that $ah_1 = bh_2$. Then $a = bh_2h_1^{-1}$ and any element ah of aH can be expressed in the form $bh_2h_1^{-1}h$. Since $h_2h_1^{-1}h \in H$, it follows that $ah \in bH$. We have therefore shown that $aH \subseteq bH$. In a similar way we can show that $bH \subseteq aH$ and therefore we conclude that $aH = bH$. One way of stating the property we have just proved is to say that two cosets either coincide or have no element in common.

Since $a \in aH$, each element of G is in some coset; thus the property just proved shows that the different cosets of H in G form a partition of G.

PROOF OF (ii). Since $a \in aH$ it is obvious that if $aH = bH$, then $a \in bH$. Conversely, suppose that $a \in bH$. Then $a \in aH \cap bH$, and part (i) of the lemma implies at once that $aH = bH$.

As an example of cosets, consider the symmetric group S_3 with multiplication table 6.14. We know ϵ is the identity, and it is easy to verify that $H = \{\epsilon, \alpha\}$ is a subgroup. By use of the table, we find the following cosets of H in S_3:

$$\begin{aligned} \epsilon H &= \{\epsilon, \alpha\}, & \alpha H &= \{\alpha, \epsilon\}, \\ \beta H &= \{\beta, \alpha\beta^2\}, & \alpha\beta H &= \{\alpha\beta, \beta^2\}, \\ \beta^2 H &= \{\beta^2, \alpha\beta\}, & \alpha\beta^2 H &= \{\alpha\beta^2, \beta\}. \end{aligned}$$

We see, therefore, that there are three different cosets of H in S_3, that every coset contains two elements, and that every element of S_3 is in exactly one of the cosets. As an illustration of Lemma 6.34 (ii), we observe $\beta H = \alpha\beta^2 H$ since $\alpha\beta^2 \in \beta H$, but that $\beta H \neq \alpha\beta H$ since $\alpha\beta \notin \beta H$.

So far we have used multiplication as the operation but, as usual, it is easy to make the necessary modifications if the operation is addition. In this case, a coset is of the form $a + H = \{a + h \mid h \in H\}$. As an example, let G be the additive group of the ring $\mathbf{Z}_{12}$ and H the subgroup $\{[0], [3], [6], [9]\}$. Then it may be verified that the *different* cosets of H in G are the following:

$$\begin{aligned}[0] + H &= \{[0],[3],[6],[9]\},\\ [1] + H &= \{[1],[4],[7],[10]\},\\ [2] + H &= \{[2],[5],[8],[11]\}.\end{aligned}$$

Let us now make the following definition.

6.35 Definition. If the group G has finite order and H is a subgroup of G, the number of distinct cosets of H in G is called the ***index*** of H in G. The index of H in G is denoted by $[G:H]$; the order of G is denoted by $|G|$.

If E is the subgroup of G consisting of the identity alone, then every coset of E contains exactly one element; so the number of cosets of E equals the number of elements of G. Using the notation just introduced, this may be expressed as $|G| = [G:E]$.

Although we have been using *coset* to mean *left coset*, we should perhaps point out that Exercise 7 below shows that there are the same number of right cosets as of left cosets of H in G. Accordingly, in the definition just given, it does not matter whether we think of left cosets or of right cosets.

We shall next prove the following theorem of Lagrange which is of fundamental importance in the study of groups of finite order.

6.36 Theorem. *Suppose that the group G has order n. If H is a subgroup of G cf order m and of index k, then $n = km$. In particular, both the order and the index of H are divisors of the order of G and we may write this as $|G| = |H| \cdot [G:H]$.*

PROOF. We first observe that every coset of H in G has exactly m elements. For if $a \in G$ and h_1, $h_2 \in H$, then $ah_1 = ah_2$ if and only if $h_1 = h_2$. Hence an arbitrary coset aH has the same number of elements as H, namely, m.

We have already shown that the distinct cosets of H in G form a partition of G. Since there are k distinct cosets and each of them contains m elements, G must contain km elements. This shows that $n = km$ and the proof is complete.

There are some interesting consequences of the fact that the order of a subgroup of a finite group is a divisor of the order of the group. First of all, since the order of an element of a group is the order of the cyclic subgroup generated by that element, we have at once the following corollary.

6.37 Corollary. *The order of an element of a group of finite order is a divisor of the order of the group.*

If the order of a group is a prime p, then every element of the group, other than the identity, must have order p. This yields the next result as follows.

6.38 Corollary. *A group of order p, where p is a prime, is a cyclic group. Moreover, every element except the identity is a generator of the group.*

If the group G has order n, and the element a of G has order m, then, by Corollary 6.37, we have $n = mk$ for some integer k. By Theorem 6.26, we know that $a^m = e$, and hence $a^n = (a^m)^k = e^k = e$. We have established the following corollary.

6.39 Corollary. *If a is an element in a group of order n, then $a^n = e$.*

Corollary 6.39 may be used to give a different proof of Fermat's Little Theorem 5.33. For a prime p, the multiplicative group of the field $\mathbf{Z}_p$ has order $p - 1$. If a is an integer not divisible by p, then the equivalence class $[a]$ of a modulo p is an element of this multiplicative group and so, by the corollary, $[a]^{p-1} = [1]$. This is equivalent to $a^{p-1} - 1$ is divisible by p as had been proved earlier. A proof of Euler's Theorem 5.32 may be given by the same reasoning.

EXERCISES

1. Exhibit all cosets of the subgroup $\{\epsilon, \alpha, \alpha^2, \alpha^3\}$ of the octic group (6.15).
2. Let G be the multiplicative group of the field $\mathbf{Z}_{19}$ and H the cyclic subgroup of G generated by the element $[8]$. Exhibit all of the cosets of H in G.
3. Exhibit all cosets of the subgroup $\{0, 4, 8, 12, 16\}$ of the additive group of $\mathbf{Z}_{20}$.
4. Prove that a group of order n has a proper subgroup if and only if n is not a prime.
5. Prove that if an abelian group G of order 6 contains an element of order 3, G must be a cyclic group.
6. Let H be a subgroup of a group G. If $a, b \in G$, let $a \sim b$ mean that $b^{-1}a \in H$. Show that "$\sim$" is an equivalence relation defined on G. If $[a]$ is the equivalence class which contains a, show that $[a] = aH$ and therefore the cosets of H in G are the equivalence classes relative to this equivalence relation.

7. Let H be a subgroup of a group G and define a mapping β of the set of left cosets of H into the set of right cosets of H as follows: $\beta(aH) = Ha^{-1}$. Prove that β is a well-defined mapping and that, in fact, it is a one-one mapping of the set of all left cosets of H onto the set of all right cosets.
8. Let G be a group of order n, and let H and K be subgroups of G of respective orders h and k. If the subgroup $D = H \cap K$ of G has order d and we set $HK = \{hk \mid h \in H, k \in K\}$, prove that HK contains exactly hk/d distinct elements. [Hint: If $\{Dk_1, Dk_2, \ldots, Dk_m\}$ is a complete set of distinct cosets of D in K, show that $\{Hk_1, Hk_2, \ldots, Hk_m\}$ is a partition of HK.]
9. Let H and K be subgroups of a group G. If $a \in G$, the set $HaK = \{hak \mid h \in H, k \in K\}$ is called a *double coset* of H and K in G.
   (i) If $a, b \in G$ and $HaK \cap HbK \neq \emptyset$, prove that $HaK = HbK$.
   (ii) If G has finite order, either prove that all double cosets of fixed subgroups H and K in G have the same number of elements or give an example to show that this need not be true.

6.6 THE SYMMETRIC GROUP S_n

We now return to a further study of permutations of a finite set $A = \{1, 2, \ldots, n\}$. We have already defined the symmetric group S_n to be the group of all permutations of A. Throughout this section the word *permutation* will mean an element of S_n for some positive integer n, and we shall sometimes find it convenient to refer to the elements of A as "symbols."

We shall first study permutations of the particular type described in the following definition.

6.40 Definition. An element α of S_n is said to be a ***cycle of length k*** if there exist distinct elements $a_1, a_2, \ldots, a_k$ $(k \geq 1)$ of A such that

$$\alpha(a_1) = a_2, \quad \alpha(a_2) = a_3, \ldots, \quad \alpha(a_{k-1}) = a_k, \quad \alpha(a_k) = a_1,$$

and $\alpha(i) = i$ for each element i of A other than $a_1, a_2, \ldots, a_k$. This cycle α may be designated by $(a_1 a_2 \cdots a_k)$.

It will be observed that a cycle of length 1 is necessarily the identity permutation. It sometimes simplifies statements to consider the identity permutation as a cycle, but we shall usually be interested in cycles of length greater than 1.

As an example of a cycle, suppose that β is the element of S_6 defined by

$$\beta(1) = 3, \quad \beta(3) = 2, \quad \beta(2) = 5, \quad \beta(5) = 6, \quad \beta(6) = 1, \quad \beta(4) = 4.$$

Then β is a cycle of length 5, and we may write $\beta = (13256)$. In a cycle, such as (13256), the symbols appearing are permuted cyclically; that is, each symbol written down maps into the next one, except that the last maps into the first. A symbol, such as 4 in this example, which is not written down is assumed to map into itself. There are other ways of writing the cycle defined above. For example, $\beta = (32561) = (25613)$, and so on. Also, in another notation introduced in Section 6.3, we have

$$\beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 2 & 4 & 6 & 1 \end{pmatrix}.$$

As further illustrations of all the various notations used, let us consider elements of S_6 and verify that

$$(146)(1345) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 6 & 5 & 4 & 1 \end{pmatrix}.$$

In the first factor 1 maps into 3, and in the second factor 3 is unchanged; hence, in the product, 1 maps into 3. The symbol 2 does not appear in either factor; hence, 2 maps into 2. In the right factor 3 maps into 4, and then in the second factor 4 maps into 6; hence, in the product, 3 maps into 6. Similarly, the other verifications are easily made.

Now let α be the cycle $(a_1 a_2 \cdots a_k)$ of S_n of length k, and let us consider the powers of α. Under the mapping α^2, we see that a_1 maps into a_3 (if $k \geq 3$), for

$$\alpha^2(a_1) = \alpha(\alpha(a_1)) = \alpha(a_2) = a_3.$$

Similarly under the mapping α^3, a_1 maps into a_4 (if $k \geq 4$), and so on. Continuing, we find that $\alpha^k(a_1) = a_1$. Since we could just as well have written $\alpha = (a_2 a_3 \cdots a_k a_1)$, a similar argument shows that $\alpha^k(a_2) = a_2$ and, in general, that $\alpha^k(a_i) = a_i$ for $i = 1, 2, \ldots, k$. It follows that $\alpha^k = \epsilon$, the identity permutation, and, moreover, k is the smallest power of α which is equal to ϵ. The following result is an immediate consequence of Theorem 6.26.

6.41 Theorem. *A cycle of length k has order k.*

Two cycles $(a_1 a_2 \cdots a_k)$ and $(b_1 b_2 \cdots b_l)$ of S_n are said to be *disjoint* if the sets $\{a_1, a_2, \ldots, a_k\}$ and $\{b_1, b_2, \ldots, b_l\}$ have no elements in common. A set of more than two cycles is said to be disjoint if each pair of them is disjoint. The next result shows why cycles play an important role in the study of permutations.

6.42 Theorem. *Every element γ of S_n that is not itself a cycle is expressible as a product of disjoint cycles.*

Before considering the proof, let us look at an example. Suppose that

$$\gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 4 & 2 & 6 & 5 \end{pmatrix},$$

and let us start with any symbol which does not map into itself, for example, the symbol 1. We see $\gamma(1) = 3$, $\gamma(3) = 4$, $\gamma(4) = 2$, and $\gamma(2) = 1$. Next take any symbol that has not yet been used and that does not map into itself—for example, 5. Then $\gamma(5) = 6$ and $\gamma(6) = 5$. Since all symbols have been used, we see that $\gamma = (1342)(56)$.

PROOF. The proof in the general case follows the same pattern as in this example. Since the identity permutation is a cycle (of length 1), we assume that γ is not the identity. Start with any symbol a_1 such that $\gamma(a_1) \neq a_1$, and suppose that $\gamma(a_1) = a_2, \gamma(a_2) = a_3, \gamma(a_3) = a_4$ and so on until we come to the point where, say $\gamma(a_k)$ equals one of the symbols $a_1, a_2, \ldots, a_{k-1}$ already used. Then we must have $\gamma(a_k) = a_1$ since every other one of these symbols is known to be the image of some symbol under the one-one mapping γ. Thus γ has the same effect on the symbols $a_1, a_2, \ldots, a_k$ as the cycle $(a_1 a_2 \cdots a_k)$, but of course γ may move other symbols whereas the cycle leaves other symbols unchanged. Suppose there is some symbol b_1 which is not equal to any of the a_i and for which $\gamma(b_1) \neq b_1$. Then proceed as above and obtain a cycle $(b_1 b_2 \cdots b_h)$. If all the symbols that do not map to themselves under γ have been used, then we have

$$\gamma = (a_1 a_2 \cdots a_k)(b_1 b_2 \cdots b_h).$$

If there is another symbol c_1 such that $\gamma(c_1) \neq c_1$, we can similarly obtain another cycle. This process can be continued until all the symbols which are moved by γ have been used and appear in some cycle. A more formal proof can be completed by using mathematical induction.

Transpositions

The cycles of length two are of special interest, and we make the following definition.

6.43 Definition. A cycle of length 2 is called ***transposition.***

A transposition (ij) merely interchanges the symbols i and j, and leaves the other symbols unchanged. Since $(ij)(ij) = \epsilon$, it follows that *a transposition is its own inverse.*

It is quite easy to show that every cycle of length more than 2 can be expressed as a product of transpositions. In fact, this result follows from the observation that

$$(a_1a_2\cdots a_k) = (a_{k-1}a_k)(a_{k-2}a_k)\cdots(a_2a_k)(a_1a_k)$$

which can be verified by direct calculation. In view of Theorem 6.42, it follows immediately that *every* permutation can be expressed as a product of transpositions. However, it is easy to verify that there is more than one way to express a permutation as such a product. As examples, we see that

$$(1234) = (34)(24)(14) = (14)(12)(32) = (14)(34)(12)(42)(34),$$
$$(123)(45) = (13)(12)(45) = (45)(23)(13) = (34)(35)(24)(14)(34),$$

and so on. Since $(ij)(ij) = \epsilon$, we can insert as many such pairs of identical transpositions as we wish. Clearly, then, a permutation can be expressed as a product of transpositions in many different ways. In spite of this apparent randomness, there is something which remains constant in these representations. We shall prove below that the number of transpositions appearing in the product for a given permutation will always be even or always be odd. As a first step toward a proof of this fact, we establish a lemma.

6.44 Lemma. *The identity ϵ of S_n cannot be expressed as a product of an odd number of transpositions.*

PROOF. We assume the lemma is false and will reach an impossible situation. Suppose we have a representation

6.45
$$\epsilon = \tau_1\tau_2\cdots\tau_k,$$

in which each τ_i is a transposition and k is odd. From all such possible products, assume that we have selected one in which k is as small as possible. Since k is odd, $k \neq 0$ and clearly $k \neq 1$, so $k \geq 3$. Assume $\tau_k = (pq)$; suppose further that from all possible products of k transpositions of the form 6.45 that we have selected one in which the symbol p appears in the least possible number of transpositions. Since the product equals the identity, p must appear in some transposition τ_j with $j \neq k$. If this were not the case, then the product on the right of 6.45 would move q to p, and so the product could not equal the identity permutation. Let j be the largest index less than k such that $\tau_j(p) \neq p$, say $\tau_j = (pr)$. Thus none of the transpositions between τ_j and τ_k involve the symbol p. The next step is to arrange to have the case $j = k - 1$; that is, these two transpositions are adjacent. This will be accomplished by using the following two formulas:

6.46
(i) $(pr)(st) = (st)(pr)$ if p, r, s, t are all different;
(ii) $(pr)(rt) = (rt)(pt)$ if p, r, t are all different.

These equations show that the nearest transposition to the left of τ_k which involves p may be moved to the right of each transposition not involving p. It may be necessary to make a change in τ_j, but the total number of transpositions remains the same—namely, k, and the number of transpositions which involve p remains the same. With all this done, we may then assume that the product of the last two transpositions in 6.45 is of the form

$$\tau_{k-1}\tau_k = (pm)(pq).$$

Now if $m = q$, the product $(pq)(pq) = \epsilon$, and so we have a product of $k - 2$ transpositions equal to the identity, a conflict with the supposition that k was the fewest possible number of transpositions which could be used. Thus $m \neq q$. Now we use the equation $(pm)(pq) = (mq)(pm)$ to replace the last two terms in 6.45 and thus obtain a product in which the number of occurrences of the symbol p is smaller than the number of occurrences in the original product. Since we had assumed this number was already the least possible, we again have an impossible situation, and the lemma must be true.

The following theorem, of which the first statement has already been proved, is one of the principal theorems about permutations.

6.47 Theorem. *Every permutation α can be expressed as a product of transpositions. Moreover, if α can be expressed as a product of r transpositions and also as a product of s transpositions, then either r and s are both even or they are both odd.*

PROOF. Suppose that α is a permutation of the set $A = \{1, 2, \ldots, n\}$. Suppose, further, that

$$\alpha = \beta_1\beta_2\cdots\beta_r = \gamma_1\gamma_2\cdots\gamma_s, \tag{6.48}$$

where each β and each γ is a transposition. To establish the theorem, we need to prove that r and s are both even or that they are both odd. Note that $(ij)(ij) = \epsilon$, so that any transposition is equal to its inverse; moreover, $\alpha^{-1} = \gamma_s^{-1}\cdots\gamma_2^{-1}\gamma_1^{-1} = \gamma_s\cdots\gamma_2\gamma_1$. Thus we have

$$\epsilon = \alpha\alpha^{-1} = \beta_1\cdots\beta_r\gamma_s\cdots\gamma_1;$$

this expresses the identity as a product of $r + s$ transpositions. By Lemma 6.44, $r + s$ is an even integer; it follows that either both r and s are even, or both r and s are odd, since the sum of an even integer and an odd integer is odd.

•**6.49 Definition.** A permutation is called an ***even*** permutation or an ***odd*** permutation according as it can be expressed as a product of an even or an odd number of transpositions.

If the permutation α can be expressed as a product of k transpositions and the permutation β can be expressed as a product of l transpositions, it is obvious that $\alpha\beta$ can be expressed as a product of $k + l$ transpositions. It follows that the product of two even, or of two odd, permutations is an even permutation, whereas the product of an odd permutation and an even permutation is an odd permutation. We shall denote the set of all even permutations in S_n by A_n. If α and β are elements of A_n, then their product $\alpha\beta$ is also in A_n. Since S_n is a finite group, Theorem 6.4(b) implies that A_n is a subgroup of S_n. The group A_n is called the *alternating group on n symbols*. We now prove a theorem which determines the number of elements of A_n.

6.50 Theorem. *The alternating group A_n is a subgroup of S_n having index $[S_n:A_n] = 2$ and order $|A_n| = n!/2$.*

PROOF. The remarks before the statement of the theorem prove A_n is a subgroup of S_n. Now we show there are just two cosets of A_n in S_n. Let τ be any odd permutation—say $\tau = (12)$ to be specific. If σ is an even permutation in S_n, then $\sigma \in A_n$ and so $\sigma A_n = A_n$. If σ is an odd permutation, then $\tau\sigma$ is an even permutation and so $\tau\sigma A_n = A_n$; this implies $\sigma A_n = \tau^{-1}A_n = \tau A_n$. Hence every coset of A_n is equal to A_n or τA_n. Moreover, these two cosets are not equal because every element in τA_n is the product of an odd number of transpositions; thus the identity is not in this coset by Lemma 6.44. Hence $[S_n:A_n] = 2$ and it follows from the formula in 6.35 that $|A_n| = |S_n|/2 = n!/2$.

We conclude this section with a result about A_n which is similar to the result that every element of S_n is a product of transpositions.

6.51 Theorem. *Every cycle of length three, (abc), in S_n lies in A_n. Moreover, every element of A_n may be expressed as a product of cycles of length three.*

PROOF. The formula $(abc) = (bc)(ac)$ shows that every cycle of length three is an even permutation and so lies in A_n. Now we show that every element of A_n can be expressed as a product of cycles of length three. Any element of A_n is a product of an even number of transpositions; if the transpositions are arranged as products of pairs of transpositions $(\tau_1\tau_2)\ldots$, we see that it is sufficient to prove that the product of any two transpositions is equal to a product of cycles of length three. This can be done directly. In the following products, different letters are assumed to stand for different symbols being permuted. We have

(i) $(ab)(cd) = (abc)(bcd)$,
(ii) $(ab)(ac) = (acb)$,
(iii) $(ab)(ab) = \epsilon = (abc)^3$.

Thus any product of an even number of transpositions can also be expressed as a product of cycles of length three.

This theorem is sometimes used to show that a certain subgroup of S_n must in fact contain A_n by proving that the subgroup contains every cycle of length three.

EXERCISES

1. In each of the following, γ is an element of S_7. Express it as a product of disjoint cycles.
   (a) $\gamma(1) = 3,\ \gamma(2) = 4,\ \gamma(3) = 1,\ \gamma(4) = 7,\ \gamma(5) = 5,\ \gamma(6) = 6,\ \gamma(7) = 2.$
   (b) $\gamma(1) = 5,\ \gamma(2) = 3,\ \gamma(3) = 4,\ \gamma(4) = 7,\ \gamma(5) = 6,\ \gamma(6) = 1,\ \gamma(7) = 2.$
   (c) $\gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 1 & 2 & 6 & 7 & 5 \end{pmatrix}.$
   (d) $\gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 3 & 1 & 5 & 4 & 7 & 6 \end{pmatrix}.$
2. Express each of the following elements of S_7 as a product of disjoint cycles:
   (a) (123)(16543),
   (b) (213456)(172),
   (c) (4215)(3426)(5671),
   (d) (1234)(124)(3127)(56).
3. Verify that a cycle of length k is an even or an odd permutation according as k is odd or even, respectively.
4. Define a function $f: S_n \to \mathbf{Z}_2$, the additive group of integers modulo 2, by the rule $f(\alpha) = 0$ if α is an even permutation, $f(\alpha) = 1$ if α is an odd permutation. Show that f is a homomorphism of S_n onto $\mathbf{Z}_2$.
5. Exhibit the elements of the alternating group A_3 and of the alternating group A_4.
6. Let G be a subgroup of the symmetric group S_n, and suppose that G contains at least one odd permutation. By a suitable modification of the proof of Theorem 6.50, prove that the set of all even permutations in G is a subgroup of G, and then prove that G contains the same number of odd permutations as of even permutations.

6.7 NORMAL SUBGROUPS AND FACTOR GROUPS

If H is a subgroup of the group G and $a \in G$, the set $aH = \{ah \mid h \in H\}$, which in Section 6.5 was called a coset, we shall for the present call a *left* coset of H in G. Similarly, the set $Ha = \{ha \mid h \in H\}$ is called a *right* coset of H in G. It need not be true that a left coset aH is equal to the right coset Ha. However, we shall be interested in subgroups which do have this property, and we therefore introduce the following definition.

6.52 Definition. A subgroup K of a group G is said to be a ***normal*** (or ***invariant***) subgroup of G if and only if $aK = Ka$ for every element a of G.

We may emphasize that this definition does not state that necessarily $ak = ka$ for each $a \in G$ and $k \in K$; it merely states that the *sets* aK and Ka coincide. In particular, if $a \in G$ and $k \in K$, there must exist an element k_1 of K (not necessarily the same element k) such that $ak = k_1 a$.

Clearly, every subgroup of an abelian group is a normal subgroup. As an example of a subgroup which is not normal, consider the symmetric group S_3 with multiplication table 6.14 and the subgroup $H = \{\epsilon, \alpha\}$ of S_3. It may be verified that $\beta H = \{\beta, \alpha\beta^2\}$, whereas $H\beta = \{\beta, \alpha\beta\}$ and hence that $\beta H \neq H\beta$, so H is not a normal subgroup of S_3. However, S_3 does have the alternating group $A_3 = \{\epsilon, \beta, \beta^2\}$ as a normal subgroup. Instead of verifying this fact by direct calculation, let us prove that for each positive integer $n > 1$, *the alternating group A_n is a normal subgroup of the symmetric group S_n*. If α is an even permutation, then $\alpha \in A_n$ and $\alpha A_n = A_n\alpha = A_n$. If α is an odd permutation, the proof of Theorem 6.50 shows that αA_n is the set of all odd permutations of S_n. A similar argument will show that also $A_n\alpha$ is the set of all odd permutations of S_n. Hence for every $\alpha \in S_n$ we have $\alpha A_n = A_n\alpha$, and A_n is therefore a normal subgroup of S_n.

Now let G be an arbitrary group and K a normal subgroup of G. Since K is normal, we need not distinguish between left cosets and right cosets; so we shall again simply call them cosets and write them as left cosets.

Let G/K denote the set of all cosets of K in G. We propose to define an operation of multiplication which will make this set into a group. Accordingly, let us define

6.53 $$(aK)(bK) = (ab)K, \qquad a, b \in G.$$

In order to verify that this does define an operation on the set of all cosets, we need to show that multiplication is well-defined by this equation. That is, we must show that if $aK = a_1K$ and $bK = b_1K$, then $(ab)K = (a_1b_1)K$. By 6.33(ii), this fact can be established by showing that if $a \in a_1K$ and $b \in b_1K$, then $ab \in (a_1b_1)K$. Suppose then that $a = a_1k$ and $b = b_1k'$, where $k, k' \in K$. Thus $ab = a_1kb_1k'$ and, since K is a normal subgroup of G, there exists $k'' \in K$

such that $kb_1 = b_1k''$. Hence $ab = a_1b_1k''k'$ and it follows that $ab \in (a_1b_1)K$, as we wished to show. This proves that 6.53 does indeed define an operation of multiplication on the set of all cosets of K in G, and we proceed to prove the following theorem.

6.54 Theorem. *Let K be a normal subgroup of the group G. With respect to the multiplication 6.53 of cosets, G/K is a group. Moreover, the mapping $\theta: G \to G/K$, defined by $\theta(a) = aK$, is a homomorphism of G onto G/K, with kernel K.*

PROOF. The associative law in G/K is an almost immediate consequence of the associative law in G, and we leave this part of the proof to the reader. Now, if e is the identity of G, then since by 6.53

$$(aK)(eK) = (eK)(aK) = aK, \qquad a \in G,$$

we see that $eK = K$ is the identity of G/K. Finally, 6.53 implies that

$$(aK)(a^{-1}K) = (aa^{-1})K = eK = K$$

and, similarly, $(a^{-1}K)(aK) = K$. Hence $a^{-1}K$ is the inverse of aK, and we have proved that G/K is a group. (Cf. Exercise 14 of Section 6.3.) Furthermore, the definition of the mapping θ shows that it is a mapping of G onto G/K, and the definition of multiplication of cosets shows that

$$\theta(ab) = (ab)K = (aK)(bK) = \theta(a)\theta(b), \qquad a, b \in G,$$

and hence that θ is a homomorphism. Finally, since K is the identity of G/K, an element a of G is in $\ker \theta$ if and only if $aK = K$, that is, if and only if $a \in K$. This completes the proof of the theorem.

The group G/K constructed in this theorem is called the *factor group* of G by K. We may point out that since the elements of the factor group G/K are the distinct cosets of K in G, if G has finite order, the order of the group G/K is the index of K in G. In fact, if G has finite order, Theorem 6.36 shows that

$$\text{order of } G/K = \frac{\text{order of } G}{\text{order of } K}.$$

In the above, we have used multiplication as the operation in G. If G is abelian and the operation is considered to be addition, it is important to keep in mind that a coset is of the form $a + K$, and the multiplication 6.53 of cosets is replaced by addition of cosets defined as follows:

$$(a + K) + (b + K) = (a + b) + K, \qquad a, b \in G. \tag{6.55}$$

In this case, the identity of G is called the "zero" as usual, and the zero of the factor group G/K is the coset K.

We have shown that if K is a normal subgroup of an arbitrary group G, then there exists a homomorphism of G, with kernel K, onto the factor group G/K. We shall next prove that "essentially" all homomorphisms of G are of this type. More precisely, we shall show that the kernel of every homomorphism of G is a normal subgroup of G and that every homomorphic image of G is isomorphic to a factor group G/K for some choice of the normal subgroup K. This is the content of the following theorem.

6.56 Fundamental Theorem on Group Homomorphisms. *Let* $\phi\colon G \to H$ *be a homomorphism of the group* G *onto the group* H *with kernel* K. *Then* K *is a normal subgroup of* G, *and* H *is isomorphic to the factor group* G/K. *More precisely, the mapping* $\alpha\colon G/K \to H$ *defined by*

6.57
$$\alpha(aK) = \phi(a), \qquad a \in G,$$

is an isomorphism of G/K *onto* H.

PROOF. Let us first show that the kernel K of a homomorphism $\phi\colon G \to H$ is necessarily a normal subgroup of G. We have already proved in Theorem 6.21 that K must be a subgroup, so there only remains to prove that it is normal. If $a \in G$ and $k \in K$, then

$$\phi(aka^{-1}) = \phi(a)\phi(k)\phi(a)^{-1}.$$

But if e is the identity of G, $\phi(e)$ is the identity of H and $\phi(k) = \phi(e)$ by definition of $\ker \phi$. Thus $\phi(aka^{-1}) = \phi(e)$ and $aka^{-1} \in \ker \phi = K$. Hence $aka^{-1} = k_1$ for some element k_1 of K. It follows that $ak = k_1 a$ and this shows that $aK \subseteq Ka$. In like manner, it can be shown that $Ka \subseteq aK$, so that $aK = Ka$ and K is indeed a normal subgroup. Thus we can now speak of the group G/K.

$$\begin{tikzcd}
G \arrow[rr, "\phi"] \arrow[dr, "\theta"'] & & H \\
& G/K \arrow[ur, "\alpha"'] &
\end{tikzcd}$$

Before proving the rest of the theorem, its meaning may perhaps be clarified by reference to the accompanying diagram. Here ϕ is the given homomorphism of G onto H, $K = \ker \phi$, and $\theta\colon G \to G/K$ is the homomorphism of G onto G/K defined by $\theta(a) = aK$, $a \in G$, as in the preceding theorem. Our present theorem may then be interpreted as stating that $\phi = \alpha\theta$—that is, $\phi(a) = \alpha(\theta(a)) = \alpha(aK)$ for each $a \in G$.

Otherwise expressed, an element a of G has the same image in H no matter which of the two paths from G to H is taken.

Now let us show that a mapping α of G/K into H is well-defined by 6.57. That is, we shall show that if $aK = a_1K$, then $\phi(a) = \phi(a_1)$. If $aK = a_1K$, then $a = a_1k$ for some element k of K. Hence $\phi(a) = \phi(a_1k) = \phi(a_1)\phi(k) = \phi(a_1)$ since $k \in \ker \phi$, and therefore $\phi(k)$ is the identity of H. This proves α is well-defined.

Next, using the fact that ϕ is a homomorphism, and the definition of multiplication of cosets, we see that for any $a, b \in G$,

$$\alpha[(aK)(bK)] = \alpha[(ab)K] = \phi(ab) = \phi(a)\phi(b) = \alpha(aK)\alpha(bK).$$

This shows that α is a homomorphism and clearly it is a mapping onto H. There remains only to prove that it is an isomorphism.

Suppose that $aK \in \ker \alpha$. If e is the identity of G, then $\phi(e)$ is the identity of H and so $\alpha(aK) = \phi(e)$. The definition of α gives us $\alpha(aK) = \phi(a)$, and so $\phi(e) = \phi(a)$. However, this implies that $a \in \ker \phi = K$, so $aK \in \ker \alpha$ implies that $a \in K$ and therefore that $aK = K$. Since K is the identity of the group G/K, we conclude that $\ker \alpha$ consists only of the identity of G/K. By Theorem 6.21 it follows that α is an isomorphism, and the theorem is proved.

EXERCISES

1. If H is a subgroup of a group G and $a \in G$, prove that $aHa^{-1} = \{aha^{-1} \mid h \in H\}$ is a subgroup of G which is isomorphic to H.
2. Prove that the intersection of two or more normal subgroups of a group G is a normal subgroup of G.
3. If G is a group, prove that the set $\{a \mid a \in G, ax = xa \text{ for every } x \in G\}$ is a normal subgroup of G.
4. Verify that the subgroup $\{\epsilon, \alpha, \alpha^2, \alpha^3\}$ of the octic group (6.15) is a normal subgroup.
5. If e is the identity of G, and H and K are normal subgroups of G with $H \cap K = \{e\}$, prove that $hk = kh$ for any $h \in H$, $k \in K$. [Hint: Show that $h^{-1}k^{-1}hk \in H \cap K$.]
6. Prove that if every right coset of a subgroup H in G is also a left coset, then H is necessarily normal. [*Hint:* If $Ha = bH$, show that $bH = aH$.]
7. Let G be the cyclic group of order 20 generated by an element a, and let H be the subgroup of G generated by a^4. Write out the cosets of H in G and verify that the quotient group G/H is a cyclic group of order 4.
8. It can be shown that the set of elements $K = \{(1), (12)(34), (13)(24), (14)(23)\}$ of the alternating group A_4 on four symbols is a normal subgroup of A_4. Without calculation, explain how you know that the group A_4/K must be cyclic.

9. Prove: If there exist exactly two left cosets (or right cosets) of a subgroup H in a group G, then H is necessarily a normal subgroup of G.

10. Let **Q** be the additive group of the field of rational numbers and **Z** the additive group of the ring of integers. Show that every element of the quotient group **Q**/**Z** has finite order. Does the group **Q**/**Z** have finite order?

11. Let K be a normal subgroup of the group G, and let A be a subgroup of the quotient group G/K. Thus we may consider A to be a set of cosets of K in G. Prove that the union of these cosets is a subgroup of G.

12. Let H and K be subgroups of a group G, with K a normal subgroup of G. Prove each of the following:

    (i) $H \cap K$ is a normal subgroup of H.
    (ii) If $HK = \{hk \mid h \in H, k \in K\}$, then HK is a subgroup of G.
    (iii) K is a normal subgroup of the group HK.

13. If H and K are as in the preceding exercise, prove each of the following:

    (i) Every element of the factor group HK/K is expressible in the form hK, $h \in H$.
    (ii) The mapping $\alpha: H \rightarrow HK/K$ defined by $\alpha(h) = hK$, $h \in H$, is a homomorphism of H onto HK/K, with kernel $H \cap K$.
    (iii) The group $H/(H \cap K)$ is isomorphic to the group HK/K.

14. Let $\theta: G \rightarrow H$ be a homomorphism of the group G onto the group H, with kernel $\theta = K$. If A is a subset of G, let us use the notation introduced in Section 1.2 and write $\theta(A)$ for the set of elements of H which occur as images of elements of A under the map θ. In particular, $\theta(G) = H$ since θ is given as an onto mapping. If U is a subset of H, let us define $\theta^{-1}(U) = \{x \mid x \in G,\ \theta(x) \in U\}$.* Prove each of the following:

    (i) If A is a subgroup of G, then $\theta(A)$ is a subgroup of H.
    (ii) If U is a subgroup of H, then $\theta^{-1}(U)$ is a subgroup of G which contains K.
    (iii) The mapping $A \rightarrow \theta(A)$ is a one-one mapping of the set of all subgroups of G which contain K onto the set of all subgroups of H.

*6.8 THE 15-PUZZLE

The 15-puzzle is a children's game (for children of all ages) to which we apply the theory of the symmetric group in order to make some adult observations.

The 15-puzzle is a 4 × 4 frame holding interlocking squares, numbered from 1 to 15 with one space left vacant (Figure 14). The squares may be moved horizontally or vertically into the vacant space, but no square may be lifted out

---

* This is here a convenient notation but is not to be confused with the inverse mapping as introduced in Section 1.2. Actually, as here used, θ^{-1} is not a mapping of H into G, but it does define a mapping of the set of all subsets of H into the set of all subsets of G.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

Starting
Position (SP)

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

1	2	3	4
5	6	7	8
9	10	11	
13	14	15	12

A Simple Move

Figure 14

of its plane. A *simple move* consists of sliding one square into the blank space, thus in effect moving the blank space to an adjacent location.

The problem we shall discuss is this: Starting from a given arrangement of the numbered squares, what new arrangements can be obtained by performing a series of simple moves? Putting it another way, find a way to determine if a given arrangement can be obtained from the starting position by a sequence of simple moves. For example, we shall be able to tell rather quickly after proving Theorem 6.58 that only one of the positions in Figure 15 may be reached from the starting position by a series of simple moves.

12	14	5	7
10		3	9
8	2	1	11
6	4	15	13

Position I

8	10	7	9
6	12	5	11
4	14	3	13
2		1	15

Position II

1	2	3	4
8	7	6	5
9	10	11	12
	15	14	13

Position III

Figure 15

Now we set up the notation which permits us to solve this problem. Use the numbers 1 through 16 to indicate *locations* in the frame which holds the squares. In the starting position square 1 is in location 1, square 2 in location 2, and so on; the blank space is location 16. The location numbers remain fixed; the numbered squares in the locations may change. After a series of moves is applied, the squares will be in different locations and the new position may be viewed as a permutation of the set $\{1, 2, \ldots, 16\}$. We take the point of view that a permutation $\sigma \in S_{16}$ is applied to any arrangement A: This produces a new arrangement $\sigma(A)$ by moving the square in location i of A to the location $\sigma(i)$ in the new arrangement $\sigma(A)$ for each $i = 1, \ldots, 16$. Thus the transposition $(15, 16)$ when applied to the starting position produces the new arrangement in which the 15 square has been moved into the blank space, and the blank space has been moved into the 15 location. This position is clearly obtainable by a simple move. The transposition $(15, 16)$, when applied to the Position I arrangement, produces a new arrangement in which the two squares numbered 13 and 15 are interchanged.

We now set up a one-one correspondence of the arrangements of the squares in the frame and the elements of S_{16}. If A is any arrangement of the squares and SP is the starting position, then there is a unique permutation $\alpha \in S_{16}$ with the property $\alpha(SP) = A$. We let the arrangement A correspond to α. We now have a correspondence of arrangements with group elements. Of course, the set of arrangements does not form a group, so there is no question of isomorphism here. But there is a group lurking in the background and we shall exploit it. If we apply a series of simple moves to the starting position and thereby produce an arrangement A corresponding to the permutation α, it would be possible to perform the same series of simple moves on any arrangement—provided that the arrangement to which the moves are applied has the blank square in location 16, the same as in the starting position. The group in the background has for its elements all the series of simple moves which leaves the blank square in its initial location 16. The product of two such series of moves is "juxtaposition"—that is, perform one series of moves after the other. This gives the main idea which will be made precise in the following theorem.

6.58 Theorem. *Let H be the subset of S_{16} consisting of all elements that correspond to arrangements which are obtained from the starting position by a series of simple moves which ends with the blank square in the location 16. Then H is a subgroup of S_{16} and H consists of all the even permutations which leave 16 fixed; that is, $H = A_{15}$.*

PROOF. The proof will require several steps.

(A) H is a subgroup of S_{16}.

Every simple move corresponds to a transposition (ab) where either a or b is the location of the blank square just before the move is applied to a position. Any sequence of simple moves which is applied to a position in which the blank is at location 16 before and after the move has the form

6.59 $$\sigma = (16, x_{t-1})(x_{t-1}, x_{t-2}) \cdots (x_2, x_1)(x_1, 16).$$

Note that in this product, any two adjacent transpositions must have a common symbol; $(ab)(bc)$ occurs in the product if the blank space is moved from location c to b and then to a by this part of the product. The 16 must occur in the first and last transposition because we begin and end with the blank in location 16; also, the locations a and b must be adjacent so that a square may be moved from a to b or from b to a. This description of the elements of H makes it clear that H is a subgroup since the product of two elements of the form 6.59 is again an element of this form.

(B) $H \subseteq A_{15}$.

Every element α in H is a permutation of the form 6.59; let us prove that α is an even permutation. Since α is the product of t transpositions, it is necessary to prove that t is even. Think of σ as moving the blank space from 16 to x_1, to x_2, and so on until the blank returns to 16. Every move of the blank is left, right, up, or down. Let f, r, u, and d stand for the number of left, right, up, and down moves of the blank space. Then $t = f + r + u + d$. Since the blank returns to its original position, the number of left moves equals the number of right moves, $f = r$, and the number of up moves equals the number of down moves, $u = d$. Hence $t = 2r + 2d$, which is an even number. Thus $\sigma \in A_{16}$. However, σ is also in S_{15}, the subgroup of S_{16} consisting of all elements which leave 16 fixed. It follows that $\sigma \in A_{16} \cap S_{15} = A_{15}$.

The next steps are designed to show that $H = A_{15}$. This will be accomplished by proving that every cycle (x, y, z) of length three is in H and then invoking Theorem 6.51 to conclude that every element of A_{15} is in H.

(C) H contains the elements $\alpha, \beta, \gamma, \sigma$ defined by

$$\begin{aligned}
\alpha &= (16,15)(14,15)(14,13)(13,9)(9,5)(5,1)(1,2)(2,3)(3,4)(4,8)(8,12)(12,16)\\
&= (1,5,9,13,14,15,12,8,4,3,2),\\
\beta &= (16,15)(15,14)(14,10)(10,6)(6,7)(7,8)(8,12)(12,16)\\
&= (6,7,8,12,15,14,10),\\
\gamma &= (16,15)(15,11)(11,12)(12,16)\\
&= (11,15,12),\\
\sigma &= (16,15)(15,14)(14,10)(10,6)(6,2)(2,3)(3,4)(4,8)(8,12)(12,16)\\
&= (4,8,12,15,14,10,6,2,3).
\end{aligned}$$

Each element is presented as a product of simple moves which are applied to the starting position to achieve the indicated permutation. The effect of each move is illustrated in Figure 16.

5	1	2	3
9	6	7	4
13	10	11	8
14	15	12	

α

1	2	3	4
5	10	6	7
9	14	11	8
13	15	12	

β

1	2	3	4
5	6	7	8
9	10	15	11
13	14	12	

γ

1	6	2	3
5	10	7	4
9	14	11	8
13	15	12	

σ

Figure 16

(D) H contains every cycle (x, y, z) with $1 \leq x, y, z \leq 15$.

We have $\gamma = (11, 15, 12) \in H$. For any $\theta \in S_{15}$ we have the relation

$$\theta(x, y, z)\theta^{-1} = (\theta(x), \theta(y), \theta(z))$$

showing that the element $\theta\gamma\theta^{-1}$ is also a cycle of length three which is easily computed by applying θ to the entries of the cycle. When $\theta \in H$, then $\theta\gamma\theta^{-1} \in H$. We first use β to conclude $\beta^5\gamma\beta^{-5} = (11, 8, 7) \in H$. Note that α and σ fix both 11 and 7 so that if τ is a power of α or a power of σ, then $\tau(11, 8, 7)\tau^{-1} = (11, \tau(8), 7)$. Now observe that every x with $1 \leq x \leq 15$, $x \neq 7, 11$, has the form $\alpha^i(8)$ or $\sigma^j(8)$ for a suitable i or j. It follows that every cycle of length three having the form $(11, x, 7)$ is in H.

Now we use two elements of this form to obtain

$$(11, x, 7)(11, y, 7)(11, x, 7)^{-1} = (x, y, 11);$$

so every element of this form belongs to H. Now finally we have the product

$$(11, z, 7)(x, y, 11)(11, z, 7)^{-1} = (x, y, z)$$

is an element of H.

Every element of A_{15} is a product of cycles of length three, and every such cycle is an element of H; thus $A_{15} \subseteq H$, and in view of (B), $H = A_{15}$, and the theorem is proved.

Now let us apply this theorem to examine the Positions I, II, and III given earlier. In order to see if Position I can be obtained from the starting position by a series of simple moves, we first apply a small number of simple moves (in this case, four) to the Position I to move the blank square to location 16. The result of this is the Position I′ (Figure 17).

12	14	5	7
10	2	3	9
8	4	1	11
6	15	13	

Position I′

Figure 17

Then Position I′ corresponds to the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 11 & 6 & 7 & 10 & 3 & 13 & 4 & 9 & 8 & 5 & 12 & 1 & 15 & 2 & 14 \end{pmatrix}$$

$$= (1, 11, 12)(2, 6, 13, 15, 14)(3, 7, 4, 10, 5)(8, 9)$$

of the starting position. The first three cycles in this product lie in A_{15} but the transposition $(8, 9)$ is not in A_{15}; hence this element is not an even permutation and the theorem implies that Position I′ cannot be obtained from the starting position by a series of simple moves. Of course, this implies that Position I cannot be so obtained either since Position I is obtained from Position I′ by a series of simple moves.

We shall let the reader decide whether Position II or Position III can be obtained from the starting position.

The theory developed in this section permits one to decide if a given arrangement can be obtained from the starting position but one does not obtain a method of carrying out the solution through the theory. In fact, the proof of Theorem 6.58 shows that any possible arrangement can be obtained by repeatedly applying the moves $\alpha, \beta, \gamma, \sigma$ in some order. In fact, this would not be a practical method to solve the puzzle since there are much shorter moves to carry out some of the required steps. Interested readers wishing to solve the 15-puzzle should use the theory to be reassured that a given arrangement is possible and then seek a solution by trial and error.

EXERCISES

1. Determine which of the Positions II and III can be obtained from the starting position by a series of simple moves.
2. Show that A_{15} is the smallest subgroup of S_{15} which contains α, β, and γ defined in (C); in other words, the element σ was not required to complete the proof of the theorem.
3. Prove the general result based upon the proof of Theorem 6.58: If N is a normal subgroup of S_n and if N contains a cycle (x, y, z), then $N = A_n$ or $N = S_n$.
4. Develop a theory of the 8-puzzle which consists of a 3×3 frame of squares numbered from 1 to 8, one space left vacant, and give a method to decide which arrangements can be obtained from a given starting position by a sequence of simple moves.

Chapter 7

Finite Groups

A great deal more is known about finite groups than is known about groups in general. One important reason for this is the ability to use certain types of counting arguments that are not available when working with infinite groups. The usual method is to construct some finite set closely related to the group in question and count elements that have special properties. This theme will be exploited throughout this chapter. One of the main results to be proved is Sylow's Theorem which asserts the existence of subgroups of prime power order in a finite group.

7.1 GROUPS ACTING ON SETS

Let G be a group and X a set. We say G *acts on* X if for each element $g \in G$ there is assigned a permutation θ_g of X in such a way that the equation $\theta_g\theta_h = \theta_{gh}$ holds for all elements $g, h \in G$. Another way to state this is that the mapping $g \to \theta_g$ from G to the symmetric group of all permutations of X is a homomorphism. If the mapping θ is one-one, then G is isomorphic to a subgroup of the group of all permutations on X; this will not always be the case, however. In most cases, the map θ is not written explicitly; an abbreviated notation in which we write gx or $g(x)$ as an abbreviated version of $\theta_g(x)$, for $g \in G$, $x \in X$, is used frequently. The condition required by the definition of an action of G upon X takes a simpler form in this notation; that is, we require $g(hx) = (gh)x$ or $g(h(x)) = (gh)(x)$ for all $g, h \in G$ and $x \in X$. Since θ is a homomorphism, θ_e must be the identity permutation on X if e is the identity of G. This means $x = \theta_e(x)$ or $x = ex = e(x)$ for all $x \in X$.

Example 1: Let $G = S_n$ and $X = \{1, 2, \ldots, n\}$ and the action of an element $g \in G$ upon the element $x \in X$ is the usual action of a permutation $x \to g(x)$.

Example 2: Let $G = S_n$ and let X be the set of all ordered pairs (i, j) with $1 \le i, j \le n$. The action of an element $g \in G$ upon the element (i, j) of X is defined by $g(i, j) = (g(i), g(j))$. Using this definition, it is easy to verify that $(gh)(i, j) = g[h(i, j)]$ since both sides are equal to $(gh(i), gh(j))$. Thus S_n acts upon the set of ordered pairs.

Example 3: Let G be any group and let $X = G$. For each $g \in G$ let θ_g be the function defined on X by the rule $\theta_g(x) = gxg^{-1}$. Note that since $X = G$, the multiplication of the elements g and x and g^{-1} is the group operation. Let us verify that θ_g is a permutation of X. First we show θ_g is one-to-one; if $\theta_g(x) = \theta_g(y)$, then $gxg^{-1} = gyg^{-1}$. Multiply this equation on the left by g^{-1} and on the right by g to conclude that $x = y$; thus θ_g is one-to-one. To show that θ_g is onto, select any z in X; then $\theta_g(g^{-1}zg) = g(g^{-1}zg)g^{-1} = z$, and this shows θ_g is an onto map. In order to show that this gives an action of G on X we must verify $\theta_g\theta_h = \theta_{gh}$. We show these functions are equal by verifying that each has the same effect upon every element of X:

$$\theta_g\theta_h(x) = \theta_g(hxh^{-1}) = g(hxh^{-1})g^{-1} = (gh)x(gh)^{-1} = \theta_{gh}(x),$$

which proves the assertion.

This is a very important example in the study of groups. We call this action of G upon G the *conjugation action*. An element gxg^{-1} is called a *conjugate* of x. There is another important property of this action. For each $g \in G$, the map θ_g is not only a permutation of the set G but it is in fact an isomorphism of the group G with itself; that is, θ_g preserves the multiplication in G. This is verified directly from the definition:

$$\theta_g(xy) = gxyg^{-1} = gxexg^{-1} = gxg^{-1} \cdot gyg^{-1} = \theta_g(x)\theta_g(y).$$

Example 4: Let G be any group and let X be the set of all subgroups of G. For each $g \in G$ and subgroup H of G, define

$$\gamma_g(H) = gHg^{-1} = \{ghg^{-1} : h \in H\}.$$

It is necessary to verify that this does define an action of G upon X. Let θ_g denote the conjugation action defined in the previous example. Then $\gamma_g(H)$ is the set of all $\theta_g(h)$ with $h \in H$. Since we saw that θ_g is an isomorphism of G with itself, it follows that θ_g maps a subgroup onto a subgroup; that is, $\gamma_g(H) = \theta_g(H)$ is a subgroup of G. Now just as in Example 3, except that subgroups replace elements, one shows that γ_g is a permutation of X and that $\gamma_g\gamma_h = \gamma_{gh}$.

By analogy with the terminology introduced just above for elements, we call gHg^{-1} a *conjugate* of the subgroup H if H is a subgroup of G and $g \in G$. The conjugate gHg^{-1} is a subgroup isomorphic to H.

Example 5: Let G be any group and H a fixed subgroup of G. For X take the set of all cosets of H, $X = \{aH \mid a \in G\}$. For each $g \in G$, define θ_g by the rule $\theta_g(aH) = gaH$. One may now show that θ_g is a permutation of X and that this defines an action of G on X.

Now that we have several examples of groups acting on sets, it is necessary to develop some properties that will be applied later in some specific situations. We assume G is a group acting on the set X.

Define a relation on X by the rule $x \sim y$ if there is some $g \in G$ for which $g(x) = y$. Let us verify that this relation is in fact an equivalence relation. Since $e(x) = x$, it follows that $x \sim x$; if $g(x) = y$, then $x = g^{-1}(y)$ and so $x \sim y$ implies $y \sim x$; finally, if $g(x) = y$ and $h(y) = z$, then $(hg)(x) = z$, and it follows that if $x \sim y$ and $y \sim z$, then $x \sim z$. The equivalence classes defined by this relation are:

7.1 $$[x] = \{y \mid y \in X \text{ and } y = g(x) \text{ for some } g \in G\}.$$

The equivalence classes are certain subsets of X defined by reference to a certain action of the group G upon X. It may be convenient at times to emphasize the role of G in this definition, so we may sometimes call the set 7.1 a *G-equivalence class*. Next we use the action of G upon X to define certain subgroups of G. For any $x \in X$ let

7.2 $$G_x = \{g \mid g \in G \text{ and } g(x) = x\}.$$

We may sometimes refer to G_x as the subgroup of G *fixing* x since it consists of all the elements of G which do not move the point x. Of course, it must be verified that G_x is indeed a subgroup. If $g, h \in G_x$, then

$$gh(x) = g(h(x)) = g(x) = x$$

so that G_x is closed under multiplication. Furthermore $h(x) = x$ implies $x = h^{-1}(x)$, so that h^{-1} is in G_x. Thus by Theorem 6.4, we see that G_x is a subgroup of G.

Now we may give our first of several counting results.

7.3 Theorem. *Let G be a group which acts on the finite set X. For each $x \in X$, the number of elements in the equivalence class $[x]$ defined by* 7.1 *is equal to the index $[G{:}G_x]$.*

PROOF. The proof of this theorem is carried out by establishing a one-one correspondence between the set of all the cosets of G_x in G and the elements in $[x]$. To establish the correspondence, we define a function ψ from the set of cosets of G_x to the elements in $[x]$ by the rule

$$\psi(gG_x) = g(x).$$

It is necessary to first show that ψ is well-defined. Suppose that $gG_x = hG_x$ and so $g = hu$ for some $u \in G_x$. Then

$$\psi(gG_x) = g(x) = hu(x) = h(u(x)) = h(x) = \psi(hG_x),$$

and so the map is well-defined. It is clearly onto $[x]$ since this set consists of all elements $g(x)$ for $g \in G$ by Definition 7.1. Finally it is necessary to show that ψ is one-one. Suppose that $\psi(gG_x) = \psi(hG_x)$ so then $g(x) = h(x)$. It follows that $x = g^{-1}h(x)$ so that the element $v = g^{-1}h$ is in G_x, and so $gv = h$ which means $h \in gG_x$. It follows that $gG_x = hG_x$ and that ψ is one-one. In view of the correspondence between cosets and the elements of $[x]$, it follows that the number of cosets, namely $[G:G_x]$, equals the number of elements in $[x]$ and this proves the theorem.

For emphasis, we state an immediate consequence of the preceding theorem.

7.4 Corollary. *If the finite group G acts upon the finite set X, then the number of elements in a G-equivalence class is a divisor of the order of G.*

We now have a relation between the number of elements in a G-equivalence class in X and the index of a subgroup of G. The next step is to "count" the elements in X using this information. Let

7.5
$$X = [x_1] \cup [x_2] \cup \cdots \cup [x_r]$$

be the partition of X produced by the equivalence relation. Since every element of X lies in one, and only one, of the sets $[x_i]$, the number of elements in X is the sum of the numbers of elements in the sets $[x_i]$. Let $m_i = [G:G_{x_i}]$, so that m_i is the number of elements in the class $[x_i]$. If we let $|X|$ denote the number of elements in X, it follows from 7.5 that

7.6
$$|X| = m_1 + m_2 + \cdots + m_r.$$

This gives a formula for the number of elements in X as a sum of certain divisors of $|G|$.

Let us give an illustration of how these formulas may be used to give purely group theoretic information.

7.7 Theorem. *Let G be a finite group and let A and B be subgroups of G. Then the number of cosets of B which have the form aB, a ∈ A equals the index [A:A ∩ B] of the subgroup A ∩ B in A.*

PROOF. Let X be the set of all cosets of B which have the form aB with $a \in A$. Our goal is to count the number of elements of X and express this number in terms of an index. Let the group A act upon X by the rule $g(aB) = (ga)B$, for $g, a \in A$. It is almost immediate that this does define an action of A upon X. There is only one A-equivalence class in this case; every element of X is A-equivalent to B. Hence the number of elements in X equals the number of elements in the equivalence class $[B]$ which, by Theorem 7.3, equals the index in A of the subgroup which fixes B. An element $a \in A$ fixes B if and only if $B = aB$; this holds if and only if $a \in B$. Since a already is in A, this holds if and only if $a \in A \cap B$. Thus X has $[A : A \cap B]$ elements as required.

EXERCISES

1. Let G be the cyclic subgroup of S_5 consisting of the powers of $\alpha = (123)(45)$. Then G has order six and operates on $X = \{1, 2, 3, 4, 5\}$. Explicitly write down the G-equivalence classes and from each equivalence class, select one element x, and determine G_x. Verify the Equation 7.6 for this example.
2. Let $G = S_3$ be the symmetric group of all permutations of $\{1, 2, 3\}$. Let X be the set of all ordered pairs (i, j) with $i, j \in \{1, 2, 3\}$. Then G acts upon X as in Example 2. Explicitly give the G-equivalence classes of X and obtain the decomposition 7.5.
3. Let G be the subgroup $\{\epsilon, \beta, \beta^2\}$ of S_3, where β is the cycle of length three, $\beta = (123)$. Repeat Exercise 2 using the X given there and this G in place of S_3.
4. Let G be a finite group which has subgroups A and B of orders m and n, respectively. If the intersection $A \cap B$ has d elements, prove that the number of distinct products ab, with $a \in A$, $b \in B$, is mn/d.
5. Let $\mathbf{Z}_3$ be the field with three elements and let G be the group of all two by two matrices of the form

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad a, b, c, d \in \mathbf{Z}_3,\ ad - bc \neq 0,$$

and let X be the set of all "columns" $z = \begin{bmatrix} u \\ v \end{bmatrix}$ with u, v any elements of $\mathbf{Z}_3$. Define an operation of G upon X by the rule

$$gz = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} au + bv \\ cu + dv \end{bmatrix}.$$

   (i) Verify that this does define an action of G upon X.
   (ii) Determine the G-equivalence classes in X and for each class $[u]$, determine the subgroup G_u for one element of $[u]$.
   (iii) In this example G has order 48 and X has 9 elements. Verify Equation 7.6 for this example.
6. Let G be a finite group which acts upon the set X and let $x, y \in X$ be elements in the same G-equivalence class. Prove that $|G_x| = |G_y|$. In fact, more is true: If $g \in G$ is such that $g(x) = y$ then $gG_xg^{-1} = G_y$.
7. Let G be a group of order 7 and let X be a set with 5 elements. Prove that the only action of G upon X is the one defined by $g(x) = x$ for all $g \in G$, all $x \in X$.
8. Generalize Exercise 7 to the case in which G is a group of prime order p, and X is a set with fewer than p elements. The conclusion is the same; the only action of G upon X is the trivial one, $g(x) = x$ for all g and x.

9. Let G be a finite group and H a subgroup of G with $[G:H] = r$. Suppose the order of G is larger than $r!$, the order of S_r. Prove that G has a normal subgroup N with $N \neq \{e\}$, and $N \neq G$. [Use the action of G on the cosets of H as described in Example 5 to obtain a homomorphism of G into S_r. Consideration of the orders shows this cannot be one-one. Make use of Theorem 6.56.]

7.2 GROUPS OF PRIME POWER ORDER

If a finite group G has order p^n, for some prime number p, then every divisor of the order of G is either equal to 1 or is itself divisible by p. This simple fact has surprising consequences which will be explored in this section.

7.8 Theorem. *If p is a prime number and G is a finite group of order p^n for a positive integer n and if G acts on a finite set X, then either*

(*i*) *the number of elements in X is divisible by p,*

or

(*ii*) *there is some $x \in X$ such that $g(x) = x$ for every $g \in G$.*

PROOF. Suppose that condition (ii) does not hold. We must prove that condition (i) does hold. For each $x \in X$, there is at least one element $g \in G$ such that $g(x) \neq x$, for otherwise statement (ii) would hold. Thus $G_x \neq G$ and so $[G:G_x] = p^s$ for some $s > 0$. In particular, this means that the number of elements in the G-equivalence class of x is divisible by p. Now we refer to Equations 7.5 and 7.6 to conclude that $|X|$ is divisible by p, as required by statement (i).

When G acts on the set X, a point $x \in X$ with the property $g(x) = x$ for every $g \in G$ is called a *fixed point* or a *G-fixed point*. The theorem states that when G has prime power order, then any set on which it acts must either have a fixed point or the number of elements in the set is a multiple of p. We may improve this theorem by giving some information about the number of fixed points.

7.9 Theorem. *Let p be a prime number and G a group of order p^n. Suppose that G acts on the finite set X and that the number of G-fixed points is f. Then $|X| \equiv f \pmod{p}$.*

PROOF. The element $x \in X$ is a fixed point if and only if the equivalence class $[x]$ has exactly one element. All classes containing more than one element must contain p^s elements for various integers $s > 0$ because the number of elements in such a class is a divisor of the order of G—that is, a divisor of p^n. When we count the number of elements in X by using 7.6, we see that $|X|$ is a sum of f 1's and other numbers which are powers of $p > 1$. It follows that $|X| \equiv f \pmod p$.

Let us give an application of this result to prove a fact about groups of prime power order.

7.10 Theorem. *Let p be a prime and G a group of order $p^n > 1$. Then there is an element x in G with $x \neq e$, the identity, and $xg = gx$ for every $g \in G$.*

PROOF. Let G act upon the set G by conjugation (see Example 3 in Section 7.1) so that for each $g \in G$ the action of g is $g(x) = gxg^{-1}$. An element x is a fixed point of G if $x = gxg^{-1}$ for each $g \in G$. This is equivalent to the condition $xg = gx$ for all $g \in G$. There is certainly at least one fixed point, namely, the identity e of G. So if f is the number of fixed points we know $f > 0$. By Theorem 7.9 we also know $f \equiv |G| \equiv 0$ (*mod p*). This means that p divides f and so $f \geq p$; in particular, there is a fixed point x other than the identity. As already pointed out above, this is all that is required to complete the proof of the theorem.

For any group G, the *center of* G is the set

$$C(G) = \{x \mid x \in G, xg = gx \text{ for all } g \in G\}.$$

Using this terminology, Theorem 7.10 says that the center of a group having order $p^n > 1$ contains a nonidentity element.

We know already that a group of order p, p a prime, must be cyclic. We now use this last result to prove something about groups of order p^2.

7.11 Theorem. *If G is a group of order p^2, p a prime, then G is abelian.*

PROOF. If G contains an element of order p^2, then G is cyclic, and hence is abelian. Suppose that G does not have any element of order p^2. Then by Lagrange's Theorem, every nonidentity element of G has order p. Let x be a nonidentity element in G which is promised by the conclusion of Theorem 7.10. Since x has order p, there must be some element y in G which is not a power of x, and in addition y has order p. We have $xy = yx$ by the choice of x. Also the number of elements of the form $x^i y^j$ must be p^2. There are several ways to verify this point. Here is one such way. Let A be the cyclic group generated by x, and B the cyclic group generated by

y. Then $A \cap B = \{e\}$ because $A \neq B$ and both are groups of prime order; hence the subgroup $A \cap B$ of A must have order 1. Now by Theorem 7.7, there are p cosets x^iB, and so there are p^2 elements of the form x^iy^j. Since G has order p^2, every element of G is equal to a power of x times a power of y. Now using the fact that $xy = yx$, it is easy to verify that $(x^iy^j)(x^ry^s) = (x^ry^s)(x^iy^j)$ for all integers i, j, r, s. Hence the commutative law holds for every pair of elements of G, and G is abelian.

This shows that groups of orders p or p^2 are abelian; there are nonabelian groups of order p^3. In fact, the octic group with multiplication table 6.15 has order $8 = 2^3$ and is a nonabelian group.

EXERCISES

1. Find the center of the octic group (6.15).
2. Let $\alpha = (123)(456)$ and $\beta = (347)$ be permutations in S_7. Both α and β have order three. Use Theorem 7.8 to reason that α and β are not contained in any subgroup of order 3^n in S_7. [This can be done in other ways but the use of Theorem 7.8 does not require any multiplication of permutations.]
3. Let G be a group of order p^n such that the center of G has order at least p^{n-1}. Use the ideas from the proof of Theorem 7.11 to show that G is abelian and so the center of G equals G.
4. Suppose that a group contains the elements g, h, and gh which all have order two. Show that $\{e, g, h, gh\}$ is an abelian group of order four.

7.3 SYLOW'S THEOREM

In this section we shall apply the results of the previous sections to prove the existence of subgroups of order p^n in a finite group whose order is divisible by p^n, where p is a prime number. The first step in the proof of this theorem is called Cauchy's Theorem; it has been proved in many different ways. The one we give here, based upon an idea of J. McKay, is perhaps the most elegant and the shortest.

7.12 Cauchy's Theorem. *If H is a finite group and p is a prime divisor of the order of H, then H contains an element of order p.*

PROOF. We use the elements of H to construct a set X, and then define an action of a cyclic group of order p upon X and then apply Theorem 7.9 to obtain the desired conclusion.

Let

$$X = \{(h_1, h_2, \ldots, h_p) \mid h_i \in H \text{ and } h_1h_2\cdots h_p = e\}.$$

We first count the number of elements of X. The first $p-1$ entries in an element of X, $h_1, \ldots, h_{p-1}$, may be selected in a completely arbitrary manner from H; thus the number of choices is $|H|^{p-1}$. The last element h_p is completely determined since it is the inverse of the product $h_1 h_2 \cdots h_{p-1}$; thus $|X| = |H|^{p-1}$. Since p divides $|H|$, this number is divisible by p. Now let σ be the function defined on X which cyclically permutes the entries by the rule

$$\sigma(h_1, h_2, \ldots, h_p) = (h_p, h_1, \ldots, h_{p-1}).$$

Notice that σ maps X into itself because

$$h_p h_1 \cdots h_{p-1} = h_p(h_1 \cdots h_{p-1} h_p)h_p^{-1} = h_p(e)h_p^{-1} = e.$$

Moreover, σ^p is the identity mapping, so the group G consisting of all powers of σ is a cyclic group of order p which acts on X. Let f be the number of fixed points of this action. Since we have already seen that $|X|$ is divisible by p, Theorem 7.9 implies that f is divisible by p. Now we see that $(e, e, \ldots, e)$ is surely an element of X which is fixed by σ, so $f \neq 0$. Since f is divisible by p, $f > 1$ and there must be a fixed point other than the one with all e's as coordinates. Let $(h_1, h_2, \ldots, h_p)$ be a fixed point under σ. Then we have

$$(h_1, h_2, \ldots, h_p) = \sigma(h_1, h_2, \ldots, h_p) = (h_p, h_1, \ldots, h_{p-1})$$

from which it follows that $h_1 = h_2 = \cdots = h_p = h$. It follows that $h \neq e$ and $h_1 h_2 \cdots h_p = h^p = e$ so that h is the required element.

One conclusion of Sylow's Theorem is that whenever p^n divides the order of the finite group H, then there is a subgroup of order p^n. Cauchy's Theorem makes a start in this direction by providing a subgroup of order p. The next step is to find a method of enlarging this to a group of order p^2, then to p^3, and so on until a subgroup of order p^n is obtained. At first one might try taking a subgroup P of order p and an element x of order p with $x \notin P$ and then considering the smallest subgroup which contains both P and x. Unfortunately, this method will not always succeed as can be seen by the Exercise 3 at the end of Section 7.2. This idea will work if P and x satisfy an additional condition—namely, $xPx^{-1} = P$, as we now show.

7.13 Theorem. *Let P and A be subgroups of the finite group H. Assume $gPg^{-1} = P$ for each $g \in A$. Then the set*

$$AP = \{xy \mid x \in A, y \in P\}$$

is a subgroup of H having order $[A : A \cap P]|P|$.

PROOF. The set AP is the union of cosets of the form gP, $g \in A$ and by Theorem 7.7, the number of such cosets is $[A:A \cap P]$. Since each coset has $|P|$ elements, the set AP has $[A:A \cap P]|P|$ elements. Now we must show that AP is actually a subgroup. The condition $gPg^{-1} = P$ is equivalent to $gP = Pg$ for all $g \in A$. It then follows (see the proof of 6.53) that

$$gPhP = ghP \qquad g, h \in A.$$

We may interpret this as saying that the product of two elements—one from gP and one from hP—lies in ghP for all g, h in A. Since every element of AP lies in a coset of P, this is the same as saying that AP is closed under multiplication. Since H is finite, this implies by Theorem 6.4 that AP is a subgroup. This proves the theorem.

7.14 Corollary. *Let A and P be subgroups of the finite group H. Suppose that each of A and P have order a power of the prime p and in addition that $gPg^{-1} = P$ for every $g \in A$. The AP is a subgroup of H which has p power order. Moreover, if A is not contained in P, then AP is larger than P.*

PROOF. The Theorem 7.13 implies that AP is a subgroup of order $[A:A \cap P]|P|$. Since $|A|$ and $|P|$ are powers of p, the order of AP is also a power of p. Thus AP is a group of order a power of p contained in H. This group will equal P only if $gP = P$ for every $g \in A$; this is the same as saying every element of A is already in P. Thus whenever A is not contained in P, the group AP is a larger group than P.

This construction provides a critical step in the proof of Sylow's Theorem which we shall present in a moment. First, we introduce the following idea. Let G be any group and A a subgroup of G.

The *normalizer of A in G* is the set

$$N_G(A) = \{g \mid g \in G, gAg^{-1} = A\}.$$

This set is a subgroup in which A is contained as a normal subgroup. This can be proved directly but, in fact, it follows from facts already proved. If we let G operate upon the set of all subgroups of G as in Example 4 of Section 7.1, then $N_G(A)$ is the subset of G which fixes A, that is, G_A in earlier notation, and this has been shown to be a subgroup.

7.15 Sylow's Theorem. *Let H be a finite group of order $p^k m$, with p a prime not dividing the integer m, and $k \geq 1$. Then:*

(i) H contains a subgroup of order p^i for each $i = 1, 2, \ldots, k$.

(*ii*) *For each i with* $1 \le i < k$, *a subgroup of order* p^i *is contained in a subgroup of order* p^{i+1}.

(*iii*) *If P and Q are subgroups of H each having order* p^k, *then there is some* $h \in H$ *with* $P = hQh^{-1}$.

(*iv*) *The number of subgroups in H having order* p^k *is an integer of the form* $1 + pt$ *for some t and* $1 + pt$ *is a divisor of m.*

PROOF. We know by Cauchy's Theorem that there is a subgroup of order p in H. Suppose that U is a subgroup of order p^i, $1 \le i < k$. We shall prove that there is a subgroup M of order p^{i+1} which contains U. Let X be the set of all cosets yU, $y \in H$. Then X has $[H:U] = p^{k-i}m$ elements and, since $i < k$, the number of elements in X is divisible by p. Let U operate upon X by the rule $u(yU) = (uy)U$ and let f be the number of U-fixed points in X. Then $f > 0$ because the coset $U = uU$ for every $u \in U$, so U is a fixed point. It follows from Theorem 7.9 that p divides f. Next we show that the set of elements of H which lies in the cosets fixed by U forms a subgroup; in fact, the subgroup is $N_H(U)$. If gU is fixed by U, then $ugU = gU$ or $g^{-1}ug \in U$ for every $u \in U$. This says that g^{-1} is in $N_H(U)$; since the normalizer is a subgroup, $g \in N_H(U)$. Conversely, if $g \in N_H(U)$, then $UgU = gU$ and so gU is a fixed point. Thus $f = [N_H(U):U]$ and this number is divisible by p. Now consider the factor group $N_H(U)/U$, the group whose elements are the cosets of U in $N_H(U)$. This is a group having order divisible by p and so by Cauchy's Theorem there is a subgroup of order p. Such a subgroup must be cyclic, so let xU be a generator. Then x^p is an element of U and so it has order a power of p. Let Q be the cyclic subgroup of H with generator x. Since x is in $N_H(U)$, Q is in $N_H(U)$. Thus we have two groups Q and U of p power orders, $gUg^{-1} = U$ for every $g \in Q$, and $[Q: Q \cap U] = p$. Apply Corollary 7.14 to conclude QU is a subgroup of order p^{i+1}. This proves (ii), and by repeating the same line of reasoning, we eventually reach a subgroup of order p^k, which proves (i).

Now we begin the proof of (iii). Let P and Q be subgroups of H each having order p^k. Let $X = \{gQg^{-1} \mid g \in H\}$; this is the set of all subgroups conjugate to Q in H. We first want to count the number of elements in X. Let Q operate on X by conjugation; that is, for $x \in Q$ and $S \in X$, the action of x upon S is $x(S) = xSx^{-1}$. It is readily verified that this does define an action of Q on X. There is a Q-fixed point of this action; namely, $Q \in X$ and $xQx^{-1} = Q$ for every $x \in Q$. There is no other fixed point, however; suppose S is in X and $xSx^{-1} = S$ for all x in Q. Then Corollary 7.14 implies QS is a group of p power order containing Q. Since $|Q| = p^k$, and no power of p larger than p^k divides $|H|$, it follows that there is no subgroup of order p^s with $s > k$. In particular, $QS = Q$ and so $S \subseteq Q$. Since $S = gQg^{-1}$, for some $g \in H$, it follows that $|S| = |Q| = p^k$ and so $S = Q$. So the number of Q fixed points is $f = 1$ and the number of elements in X is $1 + pt$, for some integer t, by Theorem 7.9. Now we act upon this set X using the subgroup P. The action of an element $y \in P$ upon an element S in X is also conjugation; $y(S) = ySy^{-1}$. If there are no P-fixed points in X, then Theorem 7.8 implies that the number of elements in X is divisible by p. However, $|X| = 1 + pt$ is not divisible by p and so there is some element $S = gQg^{-1}$ in X which is fixed by P. By the same reasoning, we conclude PS is a subgroup of p power order which cannot be larger than P. Hence $P = S = gQg^{-1}$. This proves (iii) and also proves that every subgroup of order p^k is in X. Hence the number of subgroups of order p^k is $1 + pt$, which proves part of (iv). All that remains to be proved is that $1 + pt$ divides m. We use the set X one more time, but now we act on X with the group H. Since every element of X has the form gQg^{-1}, for some $g \in H$, there is only one H-equivalence class in X. The subgroup of H which fixes Q is $N_H(Q)$ and the number of elements in X, by Theorem 7.3, is $[H:N_H(Q)] = 1 + pt$. Clearly this number is a divisor of the order of H, $p^k m$; since p^k and $1 + pt$ are relatively prime, it follows that $1 + pt$ is a divisor of m. This completes the proof of Sylow's Theorem.

If H is a group of order $p^k m$, with p a prime not dividing m, then a subgroup of H which has order p^k is called a *p-Sylow subgroup of H*. The theorem says that a finite group has a p-Sylow subgroup for each prime p which divides its order and, moreover, any two p-Sylow subgroups are conjugate. In particular, this means that any two p-Sylow subgroups are isomorphic. It is not true that any two subgroups having equal p power orders are isomorphic (see the exercises at the end of this section).

EXERCISES

1. For $p = 2$ and 3, explicitly list the p-Sylow subgroups of the symmetric group S_3. For each pair P, Q, of p-Sylow subgroups, find an element τ in S_3 with the property $\tau Q \tau^{-1} = P$.
2. Let X be the set of 2-Sylow subgroups of the symmetric group S_4, and let S_4 act by conjugation on X. Since $|X| = 3$, S_4 acts as permutations of these three elements. Number the elements of X as 1, 2, 3 so the action of S_4 may be viewed as a homomorphism of S_4 into S_3. Determine the kernel and the image of this homomorphism.
3. Show that a 2-Sylow subgroup of S_4 has order 2^3 and contains two nonisomorphic subgroups of order 2^2.
4. Determine all the primes p for which the following statement is true: The p-Sylow subgroups of S_7 are isomorphic to the p-Sylow subgroups of S_8.

5. For any subgroup A of a group G, give a direct proof that the *normalizer* of A, $N_G(A) = \{g \mid g \in G, gAg^{-1} = A\}$, is a subgroup of G.

6. For any subgroup A of a group G, define the *centralizer* of A by

$$C_G(A) = \{g \mid g \in G, \quad ga = ag \text{ for every } a \in A\}.$$

Prove $C_G(A)$ is a subgroup of G contained in $N_G(A)$.

7.4 APPLICATIONS OF SYLOW'S THEOREM

We shall give only a small number of the many possible applications of Sylow's Theorem in this section. For a first application, we show how the theorem may be used in special instances to prove that a p-Sylow subgroup is a normal subgroup of a group.

If G is a finite group and P is a p-Sylow subgroup of G then gPg^{-1} is also a p-Sylow subgroup of G. If it is known in advance that G has only one p-Sylow subgroup, then it follows that $gPg^{-1} = P$ for all $g \in G$, and hence P is a normal subgroup of G. Conversely, if P is a p-Sylow subgroup of G and P is normal in G, then there is only one p-Sylow subgroup of G because any p-Sylow subgroup can be expressed as gPg^{-1} for some $g \in G$, by Theorem 7.15(iii), but $gPg^{-1} = P$ when P is normal. The numerical information contained in Theorem 7.15 can sometimes provide this information.

Consider a group G of order $|G| = 3^2 \cdot 5 = 45$. The number of 3-Sylow subgroups has the form $1 + 3t$, for some integer t, and $1 + 3t$ is a divisor of 5. Clearly the only possibility is $t = 0$, and so the number of 3-Sylow subgroups is 1; that is, the 3-Sylow subgroup is normal. Next consider the number of 5-Sylow subgroups of G. This number is $1 + 5s$, for some integer s, and is a divisor of 3^2. Again, the only possibility for s is $s = 0$. Thus in a group of order 45, both the 3-Sylow subgroup and the 5-Sylow subgroup are normal in G. It is not too difficult to show from this information that a group of order 45 must be abelian. We shall indicate the main step in this argument in Exercise 1.

Here is an application to a wide class of groups.

7.16 Theorem. *Let p and q be primes with $p < q$. If H is a group of order pq, then either H is abelian, or p divides $q - 1$.*

PROOF. Since $q > p$, the q-Sylow subgroup Q is normal; that is, the number of q-Sylow subgroups in H is $1 + qt$, for some integer t, and $1 + qt$ divides p. If $t \neq 0$, then $1 + qt > p$, so $t = 0$. Now let P be a p-Sylow subgroup. For any x in P, $xQx^{-1} = Q$ because Q is normal and so P acts as a group of permutations of the q elements of Q. If all the P-equivalence classes of this action contain only one element, then $xy = yx$ for every $x \in P$, $y \in Q$. But there are pq elements of the form xy, $x \in P$,

$y \in Q$, so it follows that $G = PQ$ and one may now easily verify that G is abelian. If there is some P-equivalence class in Q which does not have just a single element, then let y be an element in Q such that $xy \neq yx$ for some $x \in P$. Then the P-equivalence class of y consists of the p elements xyx^{-1}, $x \in P$. Since $y \neq e$, the p elements xy^ix^{-1}, $x \in P$ are all different for each $i = 1, 2, \ldots, q - 1$. Of course the equivalence class of e contains just one element, so an application of 7.6 yields $q = 1 + pr$, for some integer r; in other words, p divides $q - 1$.

It follows from this theorem that any group of order $15 = 3 \cdot 5$, $33 = 3 \cdot 11$, $35 = 5 \cdot 7$, and so on, must be abelian. It is true that whenever p divides $q - 1$, p and q primes, there is a nonabelian group of order pq. A construction of such a group is given in the exercises.

We shall give here a construction of a group of order $2 \cdot 7$; the idea may be generalized to produce a nonabelian group of order $2n$ for any positive integer n. We present it as a subgroup of the symmetric group S_7. Let

$$\alpha = (1234567),$$
$$\beta = (27)(36)(45);$$

then $\beta^2 = \alpha^7 = \epsilon$, the identity, and most importantly we have

$$\beta\alpha\beta^{-1} = (1765432) = \alpha^{-1}.$$

If we let A be the cyclic group generated by α, and B the cyclic group generated by β, we find $xAx^{-1} = A$ for every x in B. By Theorem 7.13, AB is a subgroup of S_7 having order $[B : B \cap A]|A| = 14$, inasmuch as $|A| = 7$, and $|B| = 2$. This group is called the *dihedral group of order 14*. There is a *dihedral group of order 2n* for every positive integer n which has the form AB for a cyclic subgroup of order n in S_n and a subgroup B of order two in S_n. If α is a generator of A and β a generator of B, the equation $\beta\alpha\beta^{-1} = \alpha^{-1}$ holds. The reader is invited to write down the appropriate permutations for a few choices of n, and perhaps even for the general case of arbitrary n.

EXERCISES

1. Let M and N be normal subgroups of a group G such that $M \cap N = \{e\}$. Prove that $mn = nm$ for every $n \in N$ and $m \in M$. [Hint: Show $mnm^{-1}n^{-1}$ lies in $M \cap N$.]
2. Let F be the field $\mathbf{Z}_p$, for a prime $p > 2$ and let r be any prime divisor of $p - 1$. Use Cauchy's Theorem to show that the polynomial $x^r - 1$ has exactly r roots in F and that all these roots form a cyclic group of order r.
3. Let F be the field $\mathbf{Z}_p$, for a prime $p > 2$. For each pair of elements $a, b \in F$ with $a \neq 0$, define a function $T_{a,b}: F \to F$ by the rule

$$T_{a,b}(x) = ax + b.$$

(i) Show that $T_{a,b}$ is a permutation of the set F.

(ii) Show that the set $G = \{T_{a,b} \mid a, b \in F, a \neq 0\}$ is a group under the usual composition of functions and its order is $p(p-1)$.

(iii) For any prime divisor r of $p-1$, let A be the cyclic group of all elements of F which satisfy $x^r = 1$ (see Exercise 2). Let H be the subgroup of G consisting of all elements $T_{a,b}$ with $a \in A$ and $b \in F$. Show that H is a nonabelian group of order pr.

(iv) Describe one r-Sylow subgroup of H, and determine the number of r-Sylow subgroups of H.

4. Let G be the set of all two by two matrices $g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with $a, b, c, d \in \mathbf{Z}_p$ and $ad - bc \neq 0$.

(i) Show that G is a group of order $p(p^2-1)(p-1)$. [Hint: The first row of g may be any of the $p^2 - 1$ elements $(a, b) \neq (0, 0)$ and once a and b have been selected, the second row (c, d) can be any of the $p^2 - p$ pairs other than (ua, ub) with $u \in \mathbf{Z}_p$.]

(ii) Let U be the set of all those g which have $a = d = 1$ and $c = 0$. Let L be the set of those g which have $a = d = 1$ and $b = 0$. Show that each U and L are p-Sylow subgroups of G and find an element y in G such that $yUy^{-1} = L$.

(iii) Show that there are $p + 1$ p-Sylow subgroups of G.

5. Let p be a prime and $\sigma = (123 \cdots p)$ a cycle of length p in the symmetric group S_p. Show that the cyclic group generated by σ is a p-Sylow subgroup of S_p. How many p-Sylow subgroups of S_p are there? [Hint: Count the number of cycles of length p and the number of these in each p-Sylow subgroup.]

Chapter 8

Finite Abelian Groups

The problem of defining all finite groups is a difficult and, in fact, an unsolved problem. However, in a sense to be made precise later, it is possible to determine all finite *abelian* groups. The purpose of this chapter is to prove the fundamental results in the theory of such groups.

Let G be a finite abelian group. Throughout this chapter we shall use addition as the operation in G. Of course, everything could just as well be stated in terms of multiplication as the operation. Unless otherwise explicitly stated, we shall always assume that the group G under discussion is a nonzero group, that is, that it does not consist of the identity alone.

Let us recall the following essential facts which will be used frequently in the sequel. Since G has finite order, each element a of G has finite order. If a has order n, Theorem 6.26 shows that n is the least positive integer such that $na = 0$. The order of the zero element is one, all other elements have order greater than one. If a has order n and $k \in \mathbf{Z}$, then $ka = 0$ if and only if $n \mid k$.

The proofs of some of the results in this chapter could be shortened slightly by making use of some of the results from Chapter 7 (Cauchy's Theorem, in particular). In order to keep this chapter self-contained, however, the proofs given will be independent of these earlier results.

8.1 DIRECT SUMS OF SUBGROUPS

If G_1 $G_2, \ldots, G_r$ are subgroups of the abelian group G, we define the *sum*

$$G_1 + G_2 + \cdots + G_r$$

of these subgroups to be the set of all elements of G which can be expressed in the form

$$a_1 + a_2 + \cdots + a_r, \qquad a_i \in G_i \ (i = 1, 2, \ldots, r).$$

This set is seen to be a subgroup of G, and each G_i is contained in this subgroup since the identity 0 of G is an element of each G_i. Actually, this sum is the smallest subgroup of G which contains all the subgroups G_i.

We now make the following definition.

8.1 Definition. If G_i $(i = 1, 2, \ldots, r)$ are subgroups of the abelian group G, the sum $G_1 + G_2 + \cdots + G_r$ is said to be a ***direct sum*** if and only if the following condition is satisfied:

(i) If $a_i \in G_i$ $(i = 1, 2, \ldots, r)$ such that

$$a_1 + a_2 + \cdots + a_r = 0,$$

then each $a_i = 0$.

We shall indicate that a sum $G_1 + G_2 + \cdots + G_r$ is a direct sum by writing it in the form

$$G_1 \oplus G_2 \oplus \cdots \oplus G_r.$$

It is worth pointing out that condition (i) is equivalent to the following condition:

(ii) If $a_i, b_i \in G$ $(i = 1, 2, \ldots, r)$ such that

$$a_1 + a_2 + \cdots + a_r = b_1 + b_2 + \cdots + b_r,$$

then $a_i = b_i$ $(i = 1, 2, \ldots, r)$.

The equivalence of (i) and (ii) follows readily from the fact that the equation $a_1 + a_2 + \cdots + a_r = b_1 + b_2 + \cdots + b_r$ may be written in the form

$$(a_1 - b_1) + (a_2 - b_2) + \cdots + (a_r - b_r) = 0.$$

We leave the details of the proof of the equivalence of conditions (i) and (ii) as an exercise.

The condition (ii) for a sum $G_1 + G_2 + \cdots + G_r$ to be a direct sum is often expressed by saying that the sum is direct if and only if each element of the sum is *uniquely* expressible in the form

$$a_1 + a_2 + \cdots + a_r, \qquad a_i \in G_i \ (i = 1, 2, \ldots, r).$$

If G_i has order n_i, we see that in a sum of the type just written there are n_i

choices for a_i, and the uniqueness property just mentioned shows that the order of a direct sum $G_1 \oplus G_2 \oplus \cdots \oplus G_r$ is the product $n_1 n_2 \cdots n_r$ of the orders of the respective subgroups G_i.

Since addition is a commutative operation in G, it is clear, for example, that $G_1 \oplus G_2 = G_2 \oplus G_1$. In general, the order in which the subgroups G_i are written in the symbol for their direct sum is immaterial.

As a simple illustration of a general property, suppose that G_1 and G_2 are subgroups of G such that $G = G_1 \oplus G_2$. Now if $G_1 = H_1 \oplus H_2$ and $G_2 = K_1 \oplus K_2$, where H_1 and H_2 are subgroups of G_1, and K_1 and K_2 are subgroups of G_2, then all of H_1, H_2, K_1, K_2 are subgroups of G and

$$G = H_1 \oplus H_2 \oplus K_1 \oplus K_2.$$

It will be clear that a similar result holds for any number of summands. (See Exercise 3.) This fact will be useful later on.

REMARK. To avoid any possible confusion, let us point out that if $G_1, G_2, \ldots, G_r$ are *any* additively written abelian groups (not necessarily subgroups of a given group), according to Section 6.1 the direct sum of these groups would consist of all ordered r-tuples

$$(a_1, a_2, \ldots, a_r),$$

with $a_i \in G_i$ for $i = 1, 2, \ldots, r$, and with addition defined as follows:

$$(a_1, a_2, \ldots, a_r) + (b_1, b_2, \ldots, b_r) = (a_1 + b_1, a_2 + b_2, \ldots, a_r + b_r).$$

Now if the G_i are subgroups of a group G and their sum is direct as defined in 8.1, it is not difficult to prove that the mapping

$$(a_1, a_2, \ldots, a_r) \to a_1 + a_2 + \cdots + a_r$$

is an isomorphism of the direct sum as defined in Section 6.1 onto the direct sum as defined in this section. Prove it! This fact justifies the use of the term *direct sum* in two different situations. Sometimes, the direct sum as defined in 8.1 is called an *internal* direct sum (since all groups G_i are subgroups of a given group G and therefore their direct sum is a subgroup of G), and the direct sum of Section 6.1 is called an *external* direct sum (since the G_i are arbitrary groups, not given as subgroups of some given group).

Now let G be an abelian group of order $n = p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k}$, where the p's are distinct primes, $k \geq 1$, and each $e_i \geq 1$. Thus $p_1, p_2, \ldots, p_k$ are the distinct prime divisors of n. Let $G(p_i)$ be the set of all elements of G having order a power of p_i. The order of the identity 0 of G is $1 = p_i^0$ and hence $0 \in G(p_i)$. Actually, $G(p_i)$ is a subgroup of G since one can see as follows that $G(p_i)$ is closed under addition. If $a, b \in G(p_i)$, suppose that a has order p_i^m and b has order p_i^n. If t is the larger of n and m, then $p_i^t(a+b)=0$, and the order of $a+b$ is a divisor of p_i^t; hence the order is a power of p_i. It follows that $a+b \in G(p_i)$, and by Theorem 6.4 we see that $G(p_i)$ is a subgroup of G. Our next goal is to prove the following theorem.

8.2 Theorem. *Let G be an abelian group of order n, and let $p_1, p_2, \ldots, p_k$ be the distinct prime divisors of n. If $G(p_i)$ is the subgroup of G consisting of all elements having order a power of p_i, then*

8.3
$$G = G(p_1) \oplus G(p_2) \oplus \cdots \oplus G(p_k).$$

Much later in this chapter we shall prove that no one of the subgroups $G(p_i)$ consists of the zero alone. In fact, if $p_i^{e_i}$ is the highest power of p_i which divides n, we shall show that $G(p_i)$ has order $p_i^{e_i}$.

Before proving Theorem 8.2, let us introduce some lemmas that will be helpful in carrying out the proof.

8.4 Lemma. *Suppose that the element a of an abelian group G has order dividing n. If m is an integer such that $(m,n)=1$, then $ma=0$ implies that $a=0$.*

PROOF. Since $(m,n)=1$, there exist integers x and y such that $1 = xm+yn$. Hence $a = xma + yna$. We are assuming that $ma=0$, and $na=0$ since a has order dividing n. It follows that $a=0$, as we wished to show.

8.5 Lemma. *If the element of the abelian group G has order $n=kl$ with $(k,l)=1$, then there exist elements b and c of G such that $a=b+c$, with b and c having respective orders k and l.*

PROOF. Since $(k,l)=1$, there exist integers s and t such that $1=sk+tl$. Thus we have $a = ska + tla$. Let us show that ska has order l. Clearly, $lska = sna = 0$. Moreover, if $z \in \mathbf{Z}$ such that $zska=0$, then $n \mid zsk$. But $n=kl$, so we conclude that $k \mid zsk$ or $l \mid zs$. Now the equation $1=sk+tl$ implies that $(s,l)=1$ and therefore $l \mid z$. Accordingly, we conclude that ska has order l. Similarly, tla has order k and if we set $b=tla$ and $c=ska$, we have $a=b+c$, b of order k and c of order l. This completes the proof.

We leave as an exercise the proof by induction of the following generalization of the preceding lemma.

8.6 Lemma. *If the element a of the abelian group G has order $n = n_1 n_2 \cdots n_k$, where $(n_i, n_j) = 1$ for $i \neq j$, then a can be expressed in the form*

$$a = b_1 + b_2 + \cdots + b_k,$$

where b_i has order n_i $(i = 1, 2, \ldots, k)$.

PROOF OF 8.2. Let us now return to the proof of Theorem 8.2 and show first that the sum $G(p_1) + G(p_2) + \cdots + G(p_k)$ is a direct sum. To this end, suppose that

$$\textbf{8.7} \qquad a_1 + a_2 + \cdots + a_k = 0, \qquad a_i \in G(p_i).$$

By definition of direct sum, we need to prove that each $a_i = 0$. For convenience of notation, let us concentrate on proving that $a_1 = 0$. Each a_i has order a power of p_i, so let us assume that a_i has order $p_i^{t_i}$ $(i = 1, 2, \ldots, k)$. From Equation 8.7, it now follows that

$$p_2^{t_2} p_3^{t_3} \cdots p_k^{t_k} a_1 = 0.$$

Since this coefficient of a_1 is relatively prime to the order of a_1, it follows from Lemma 8.4 that $a_1 = 0$. Similarly, each $a_i = 0$, and this proves that the sum is direct.

Clearly, $G(p_1) \oplus G(p_2) \oplus \cdots \oplus G(p_k) \subseteq G$, so we only need to obtain inclusion the other way. By Corollary 6.37, every element a of G has order a divisor of the order n of G, and therefore the order of a has no prime divisors except for some or all of the p_i $(i = 1, 2, \ldots, k)$. For convenience of notation only, suppose that the order of a contains only the prime factors $p_1, p_2, \ldots, p_u$, $u \leq k$. By Lemma 8.6, a is expressible as a sum of elements of $G(p_i)$, $i = 1, 2, \ldots, u$. In particular, every element of G is a sum of elements of some or all of the $G(p_i)$, $i = 1, 2, \ldots, k$. We therefore conclude that $G \subseteq G(p_1) \oplus G(p_2) \oplus \cdots \oplus G(p_k)$, and this completes the proof of the theorem.

8.2 CYCLIC SUBGROUPS AND BASES

We shall continue to let G be a finite abelian group. If $a \in G$, let us denote by (a) the cyclic subgroup generated by a. If a has order n, then

$$(a) = \{0, a, 2a, \ldots, (n-1)a\}.$$

8.8 Definition. If $a_1, a_2, \ldots, a_k$ are nonzero elements of G such that the sum $(a_1) + (a_2) + \cdots + (a_k)$ is direct, we say that the elements a_1, $a_2, \ldots, a_k$ are *independent* or form an *independent set*.

Suppose that a_i has order n_i $(i = 1, 2, \dots k)$. Then, by definition of direct sum, the a_i $(i = 1, 2, \dots, k)$ are independent if for integers z_i,

$$z_1a_1 + z_2a_2 + \cdots + z_ka_k = 0$$

if and only if $z_ia_i = 0$, that is, if and only if $n_i \mid z_i$ $(i = 1, 2, \dots, k)$.

Observe that a single element a of G is independent if and only if $a \neq 0$. Clearly, any nonempty subset of an independent set is also independent.

8.9 Definition. The set $\{a_1, a_2, \dots, a_r\}$ forms a *basis* of the abelian group G if and only if the elements of this set are independent and

$$G = (a_1) \oplus (a_2) \oplus \cdots \oplus (a_r).$$

Otherwise expressed, the group G has a basis if and only if it is expressible as a direct sum of a finite number of cyclic subgroups.

We may remark that if $G = H_1 \oplus H_2$, where H_1 is a subgroup of G having basis $\{b_1, b_2, \dots, b_s\}$ and H_2 is a subgroup of G having basis $\{c_1, c_2, \dots, c_t\}$, then G has a basis $\{b_1, b_2, \dots, b_s, c_1, c_2, \dots, c_t\}$. (See Exercise 6.)

One of the principal theorems which we shall prove in the next section is the following.

8.10 Theorem. *Every finite abelian group has a basis, each element of which has order a power of a prime.*

In view of our definitions, an equivalent formulation of this theorem would be the assertion that every finite abelian group can be expressed as the direct sum of cyclic subgroups, each of which has order a power of a prime.

By a generalization of the remark made above, the result of Theorem 8.2 shows that Theorem 8.10 will be true in general when we have established it for each of the groups $G(p_i)$. In the next section we study in some detail a class of groups which will include those of the form $G(p_i)$ as defined in Theorem 8.2.

EXERCISES

1. Prove the equivalence of conditions (i) and (ii) in connection with Definition 8.1.
2. Suppose that G_i $(i = 1, 2, \dots, r)$ are subgroups of the abelian group G such that the sum $G_1 + G_2 + \cdots + G_r$ is direct. If, for each i, H_i is a subgroup of G_i, prove that the sum $H_1 + H_2 + \cdots + H_r$ is direct.
3. Suppose that $G = G_1 \oplus G_2$, where $G_1 = H_1 \oplus H_2 \oplus H_3$ and $G_2 = K_1 \oplus K_2$. Prove that $G = H_1 \oplus H_2 \oplus H_3 \oplus K_1 \oplus K_2$. Choose an appropriate notation and generalize to an arbitrary finite number of summands.

4. Let G be the additive group of the ring $\mathbf{Z}_{24}$. In the notation of Theorem 8.2, determine the elements of $G(2)$ and of $G(3)$. Verify that these are subgroups of G and that $G = G(2) \oplus G(3)$, thus directly verifying Theorem 8.2 for this particular group.

5. Give an example to show that in a finite nonabelian group the elements which have order a power of some fixed prime need not be a subgroup.

6. Suppose that G_1 and G_2 are subgroups of the abelian group G such that $G = G_1 \oplus G_2$. If $\{a_1, a_2, \ldots, a_r\}$ is a basis of G_1 and $\{b_1, b_2, \ldots, b_s\}$ is a basis of G_2, prove that $\{a_1, a_2, \ldots, a_r, b_1, b_2, \ldots, b_s\}$ is a basis of G. Generalize to direct sums of an arbitrary finite number of subgroups.

7. Illustrate Theorem 8.10 by verifying that for the group G which is the additive group of the ring $\mathbf{Z}_{24}$, a basis of the required kind is $\{3, 8\}$.

8. Prove Lemma 8.6.

9. Prove that a cyclic group of order p^k, where p is a prime and $k \geq 1$, cannot be expressed as a direct sum of two nonzero subgroups. [Hint: Consider the maximal order that an element can have.]

10. If (a) is a cyclic group of order kl with $(k, l) = 1$, prove that there exist elements b and c of (a) of respective orders k and l, such that $(a) = (b) \oplus (c)$.

11. If b and c are elements of an abelian group G with orders k and l respectively, and if $(k, l) = 1$, prove that the sum $(b) + (c)$ is direct and that $(b) \oplus (c)$ is a cyclic subgroup of G of order kl.

8.3 FINITE ABELIAN p-GROUPS

Let us begin with the following definition.

8.11 Definition. Let p be a fixed prime. A group is said to be a ***p-group*** if the order of each of its elements is a power of p.

We may observe that the identity element (the zero) has order p^0. Every other element of a p-group has order p^m for some positive integer m. Thus a nonzero element a of a p-group has order p^m if and only if $p^m a = 0, p^{m-1}a \neq 0$. Moreover, if a has order p^m and the order of an element b of the p-group is less than or equal to the order of a, then $p^m b = 0$.

The main goal of this section is to prove the following special case of Theorem 8.10.

8.12 Lemma. *A finite abelian p-group has a basis.*

Throughout this section, let p be a fixed prime and G a finite abelian p-group. As a first step in the proof of Lemma 8.12, we collect a few useful facts for easy reference.

Let H be a subgroup of G and suppose that a is an element of G of order p^m. Since $p^m a = 0 \in H$, there exists a smallest positive integer z (necessarily less than or equal to p^m) such that $za \in H$. Throughout this section it will be convenient to have a distinctive name for this positive integer z. We call it the *order of a modulo H*. This terminology is explained by the fact that the order of a modulo H is the order of the coset $a + H$ in the factor group G/H, of G modulo H.

(A) *(i) If z is the order of a modulo H and $n \in \mathbf{Z}$, then $na \in H$ if and only if $z \mid n$. In particular, if z is the order of a modulo H and $za = 0$, then z is the order of a.*
(ii) If a has order p^m, and z is the order of a modulo H, then $z \mid p^m$ and therefore the order of each element of G modulo any subgroup of G is a power of p.

PROOF. To prove (i), let us use the Division Algorithm to write $n = qz + r$, where $0 \leq r < z$. Thus $na = q(za) + ra$. Now $za \in H$, so if $na \in H$, it follows that $ra \in H$. Since z is the smallest positive integer such that $za \in H$, we conclude that $r = 0$ and therefore $n = qz$. Conversely, if $z \mid n$, it is trivial that $na \in H$. The last statement of (i) follows from the observation that the given conditions imply that z and the order of a divide each other.

Part (ii) follows from (i) by observing that $p^m a = 0 \in H$, and therefore $z \mid p^m$ and z must therefore be a power of p.

(B) *Suppose that H is a subgroup of G and that $a \notin H$. If the order p^m of a is equal to the order of a modulo H, then the sum $H + (a)$ is direct.*

PROOF. To see the truth of this statement, suppose that $h + xa = 0$, where $h \in H$ and $x \in \mathbf{Z}$, and let us prove that $h = 0$ and that $xa = 0$. Now $xa \in H$ and since a has order p^m modulo H, part (i) of (A) shows that $p^m \mid x$. But this implies that $xa = 0$ since p^m is also the order of a. The equation $h + xa = 0$ then shows that $h = 0$. The sum $H + (a)$ is therefore a direct sum, as we wished to show.

The proof of Lemma 8.12 is carried out by the process of induction. We illustrate the approach by a fairly detailed account of the first two steps in this procedure, and then pass on to the general situation.

Let a_1 be an element of G of maximum order, say p^{m_1}. If it happens that $G = (a_1)$, we have found a basis $\{a_1\}$ of G consisting of the single element a_1. Suppose, then, that $G \neq (a_1)$. Since a_1 has maximum order among the elements of G, we see that $p^{m_1}c = 0$ for *every* element c of G.

Since $G \neq (a_1)$, we proceed to seek an element a_2 of G such that a_1, a_2 are independent and therefore $(a_1) \oplus (a_2) \subseteq G$.

Let b be an element of G of maximum order, say p^{m_2}, modulo (a_1). Since $p^{m_1}b = 0 \in (a_1)$, it follows that $p^{m_2} \leq p^{m_1}$, that is, $m_2 \leq m_1$. Since b has maximum order modulo (a_1), we see that $p^{m_2}c \in (a_1)$ for every element c of G. Thus if $c \in G$, there exists $y \in \mathbf{Z}$ such that

8.13
$$p^{m_2}c = ya_1.$$

We proceed to show that $p^{m_2} \mid y$. Multiplying the preceding equation by $p^{m_1 - m_2}$, we find that $p^{m_1}c = p^{m_1 - m_2}ya_1$. But $p^{m_1}c = 0$, and since a_1 has order p^{m_1}, we conclude that $p^{m_1} \mid p^{m_1 - m_2}y$, that is, that $p^{m_2} \mid y$. Thus $y = up^{m_2}$, where $u \in \mathbf{Z}$. Now applying what we have just proved to the special case in which c is an element b of maximum order modulo (a_1), we find from 8.13 that there exists $u \in \mathbf{Z}$ such that

8.14
$$p^{m_2}b = p^{m_2}ua_1.$$

We now set

8.15
$$a_2 = b - ua_1,$$

and observe that if $z \in \mathbf{Z}$, then $za_2 \in (a_1)$ if and only if $zb \in (a_1)$. Thus the order of a_2 modulo (a_1) is p^{m_2}. Now 8.14 and 8.15 imply that $p^{m_2}a_2 = 0$, and it follows from (A) (i) that the order of a_2 is p^{m_2}. We now know from (B) that the sum $(a_1) + (a_2)$ is a direct sum, and we have $G \supseteq (a_1) \oplus (a_2)$. If $G = (a_1) \oplus (a_2)$, we have exhibited a basis $\{a_1, a_2\}$. Otherwise, we can continue this process. We have just completed the case $k = 2$ of the following induction procedure.

Assume that we have found elements $a_1, a_2, \ldots, a_k$ of G of respective orders $p^{m_1}, p^{m_2}, \ldots, p^{m_k}$ such that all of the following are true:

(i) $m_1 \geq m_2 \geq \cdots \geq m_k$.
(ii) If $c \in G$, there exist integers $y_1, \ldots, y_{k-1}$ such that $p^{m_k}c = y_1a_1 + \cdots + y_{k-1}a_{k-1}$, and $p^{m_k} \mid y_i$ $(i = 1, 2, \ldots, k-1)$.
(iii) $a_1, a_2, \ldots, a_k$ are independent.

For convenience, let us set $G_{k-1} = (a_1) \oplus (a_2) \oplus \cdots \oplus (a_{k-1})$ and $G_k = (a_1) \oplus (a_2) \oplus \cdots \oplus (a_k)$. If $G \neq G_k$, we propose to find another element a_{k+1} of G such that all of the above three properties are true with k replaced by $k + 1$.

We may point out that (ii) above states not only that if $c \in G$, then $p^{m_k}c \in G_{k-1}$, but even gives some additional information in that the integers y_i are all divisible by p^{m_k}.

Let b be an element of G of maximum order, say $p^{m_{k+1}}$, modulo G_k. Thus, if $c \in G$, then $p^{m_{k+1}}c \in G_k$. Observe that since $p^{m_k}b \in G_{k-1} \subseteq G_k$, $p^{m_{k+1}} \leq p^{m_k}$ and $m_k \geq m_{k+1}$, so (i) holds for $k + 1$.

If $c \in G$, since $p^{m_{k+1}}c \in G_k$, there exist integers $z_1, \dots, z_k$ such that

8.16 $$p^{m_{k+1}}c = z_1a_1 + z_2a_2 + \cdots + z_ka_k.$$

We propose to show that $p^{m_{k+1}} \mid z_i$ for $i = 1, 2, \dots, k$. By (ii) of our induction hypothesis, there exist integers $y_1, \dots, y_{k-1}$ such that

8.17 $$p^{m_k}c = y_1a_1 + \cdots + y_{k-1}a_{k-1}, \quad p^{m_k} \mid y_i \; (i = 1, 2, \dots, k).$$

If we multiply 8.16 by $p^{m_k - m_{k+1}}$, we obtain $p^{m_k}c = p^{m_k - m_{k+1}}z_1a_1 + \cdots + p^{m_k - m_{k+1}}z_ka_k$. By equating the right sides of the two preceding equations, we obtain

$$(p^{m_k - m_{k+1}}z_1 - y_1)a_1 + \cdots + (p^{m_k - m_{k+1}}z_{k-1} - y_{k-1})a_{k-1} \\ + p^{m_k - m_{k+1}}z_ka_k = 0.$$

Since, by assumption, $a_1, a_2, \dots, a_k$ are independent, for each i the coefficient of a_i in this equation must be divisible by the order of a_i. In particular, $p^{m_k} \mid p^{m_k - m_{k+1}}z_k$ and this implies that $p^{m_{k+1}} \mid z_k$. Now for $1 \leq i < k$, we have $p^{m_i} \mid (p^{m_k - m_{k+1}}z_i - y_i)$. But $i < k$ and so $m_i \geq m_k$, and thus

$$p^{m_k} \mid (p^{m_k - m_{k+1}}z_i - y_i).$$

But as indicated in 8.17, we know that $p^{m_k} \mid y_i$. It follows that $p^{m_k} \mid p^{m_k - m_{k+1}}z_i$ and hence that $p^{m_{k+1}} \mid z_i$. Since this is true for $1 \leq i < k$, and we have already shown that $p^{m_{k+1}} \mid z_k$, we conclude that $p^{m_{k+1}}$ divides every z_i in 8.16. This establishes part (ii) of our induction statement for the case in which k is replaced by $k + 1$.

Now let us apply what we have just proved to the special case in which the element c in 8.16 is chosen to be a particular element b of maximal order $p^{m_{k+1}}$ modulo G_k. Thus, since $p^{m_{k+1}} \mid z_i$, there exist integers $u_1, \dots, u_k$ such that

8.18 $$p^{m_{k+1}}b = p^{m_{k+1}}u_1a_1 + p^{m_{k+1}}u_2a_2 + \cdots + p^{m_{k+1}}u_ka_k.$$

We next define

8.19 $$a_{k+1} = b - u_1a_1 - \cdots - u_ka_k,$$

and observe from this equation that a_{k+1} has the same order modulo G_k as does b, namely, $p^{m_{k+1}}$. The last two equations show that $p^{m_{k+1}}a_{k+1} = 0$ and it follows from (A)(i) that the order of a_{k+1} is $p^{m_{k+1}}$. By (B), the sum $G_k + (a_{k+1})$ is direct and it follows that $a_1, a_2, \dots, a_{k+1}$ are independent. We have shown that if $a_1, a_2, \dots, a_k$ satisfy the induction hypotheses (i), (ii), and (iii), and if $G \neq (a_1) \oplus \cdots \oplus (a_k)$, there exists an element a_{k+1} such that $a_1, \dots, a_{k+1}$ satisfy (i), (ii), and (iii), with k replaced by $k + 1$. In particular, $G \supseteq (a_1) \oplus \cdots \oplus (a_{k+1})$. Since G is assumed to be a *finite* p-group, these steps must come to an end, and thus for some positive integer r, there exist elements $a_1, a_2, \ldots, a_r$ such that

$$G = (a_1) \oplus (a_2) \oplus \cdots \oplus (a_r). \tag{8.20}$$

This completes the proof of Lemma 8.12, which states that every finite abelian p-group has a basis.

In proving 8.20, we have obtained the basis elements $a_1, a_2, \ldots, a_r$ such that their orders are respectively $p^{m_1}, p^{m_2}, \ldots, p^{m_r}$ with $m_1 \geq m_2 \geq \cdots \geq m_r \geq 1$. It follows from 8.20 that the order of G is the product of the orders of the cyclic groups (a_i), namely p^t, where $t = m_1 + m_2 + \cdots + m_r$. In particular, this shows that *the order of a finite abelian p-group is a power of p.*

Before returning to the study of arbitrary finite abelian groups, let us discuss the question of the uniqueness of a basis for a p-group. Clearly, our construction of a basis indicates that a basis is *not* unique since, as a simple example, a_1 might have been chosen to be any element of maximal order. However, we shall prove the following result.

8.21 Theorem. *Any two bases of a finite abelian p-group have the same number of elements. Moreover, the orders of the elements of one basis coincide, in some arrangement, with the orders of the elements of any other basis.*

PROOF. In proving this theorem, we shall assume that the p-group G has a basis $\{a_1, a_2, \ldots, a_r\}$, with a_i having order p^{m_i}; and that G also has a basis $\{b_1, b_2, \ldots, b_s\}$, with b_i having order p^{n_i}. Moreover, we assume that the notation is chosen so that $m_1 \geq m_2 \geq \cdots \geq m_r \geq 1$ and $n_1 \geq n_2 \geq \cdots \geq n_s \geq 1$. We proceed to prove that $r = s$ and that $m_i = n_i$ $(i = 1, 2, \ldots, r)$.

It will be convenient to consider subgroups pG and G_p of G, defined as follows:

$$pG = \{px \mid x \in G\},$$
$$G_p = \{x \mid x \in G, px = 0\}.$$

Thus $pG = \{0\}$ if and only if $G_p = G$.

Making use of the basis $\{a_1, a_2, \ldots, a_r\}$ of G, we leave it to the reader to verify that

$$\{p^{m_1 - 1}a_1, p^{m_2 - 1}a_2, \ldots, p^{m_r - 1}a_r\} \tag{8.22}$$

is a basis of the p-group G_p. Since each of these basis elements has order p and G_p is the direct sum of the cyclic groups generated by these elements, we conclude that G_p has order p^r. In exactly the same way, using the basis $\{b_1, b_2, \ldots, b_s\}$ of G, we see that G_p has order p^s. Hence $p^r = p^s$, and $r = s$. This completes the proof of the first statement of the theorem.

The proof of the second statement is by induction on the order of G, and we therefore assume as an induction hypothesis that the statement is true for all p-groups with order less than the order of G. We now make two cases; in the first we do not need this induction hypothesis.

CASE 1. $pG = \{0\}$. In this case, every nonzero element (in particular, every basis element) of G has order p. Hence, $m_i = n_i = 1$ $(i = 1, 2, \ldots, r)$.

CASE 2. $pG \neq \{0\}$. In this case pG is a nonzero subgroup of G. Moreover, the order of pG is less than the order of G, since G necessarily has some elements of order p. Why? Using the notation in which the order of a_i is p^{m_i}, not all m_i can equal 1. Suppose that u is a positive integer so chosen that $m_1 \geq m_2 \geq \cdots \geq m_u > m_{u+1} = \cdots = m_r = 1$. It may now be verified that pG has a basis

8.23
$$\{pa_1, \ldots, pa_u\}.$$

In like manner, making use of the other given basis $\{b_1, \ldots, b_r\}$ of G, if v is the positive integer so chosen that $n_1 \geq n_2 \geq \cdots \geq n_v > n_{v+1} = \cdots = n_r = 1$, we see that pG has a basis

8.24
$$\{pb_1, \ldots, pb_v\}.$$

Thus the p-group pG has bases 8.23 and 8.24. By the first statement of the theorem, already proved, we conclude that $u = v$. Using the fact that the order of pa_i is $p^{m_i - 1}$ and the order of pb_i is $p^{n_i - 1}$, the induction hypothesis as applied to the group pG shows that $m_i - 1 = n_i - 1$ for $i = 1, 2, \ldots, u$. Since all other m_i and n_i are equal to 1, we have that $m_i = n_i$ $(i = 1, 2, \ldots, r)$. This concludes the proof of the theorem.

8.4 THE PRINCIPAL THEOREMS FOR FINITE ABELIAN GROUPS

Henceforth we shall let G be an arbitrary finite abelian group. Let us assume that G has order n with distinct prime factors $p_1, p_2, \ldots, p_k$. Thus

8.25
$$n = p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k},$$

where $e_i > 0$ for all i. If $G(p_i)$ denotes the subgroup of G consisting of all elements of order a power of p_i, we have proved in Theorem 8.2 that

8.26
$$G = G(p_1) \oplus \cdots \oplus G(p_k).$$

Now $G(p_i)$ is a p_i-group and, as indicated shortly before the statement of Theorem 8.21, its order is a power of p_i. Moreover, from Equation 8.26, the order n of G must be the product of the orders of the groups $G(p_i)$. In view of the unique factorization of n into a product of primes, we conclude from 8.25 that the order of $G(p_i)$ must be $p_i^{e_i}$ $(i = 1, 2, \ldots, k)$. We have therefore proved the first statement of the following lemma.

8.27 Lemma. *Let G be an abelian group of order n.*

(*i*) *If p is a prime divisor of n, let p^e be the highest power of p which divides n. Then the subgroup $G(p)$ of G which consists of all elements with order a power of p has order p^e. In particular, $G(p) \neq \{0\}$.*

(*ii*) *If p is a prime divisor of n, then G contains an element of order p.*

PROOF. The proof of part (ii) follows at once from the observation that if a is a nonzero element of $G(p)$, then a has order p^t for some positive integer t. Hence $p^{t-1}a$ has order p.

The results of the preceding section show that each subgroup $G(p_i)$ occurring in 8.26 has a basis, say $\{a_{i1}, a_{i2}, \ldots, a_{ir_i}\}$, and clearly each element of this basis has order a power of p_i. Using this information, Equation 8.26 shows that G has a basis

8.28 $$\{a_{11}, a_{12}, \ldots, a_{1r_1}; a_{21}, a_{22}, \ldots, a_{2r_2}; \ldots; a_{k1}, a_{k2}, \ldots, a_{kr_k}\},$$

and each element of this basis has order a power of a prime. This result was stated as Theorem 8.10, one of our principal goals. We have therefore proved part (i) of the following fundamental theorem.

8.29 FUNDAMENTAL THEOREM ON FINITE ABELIAN GROUPS

(*i*) *Every finite abelian group G has a basis, each element of which has order a power of a prime.*

(*ii*) *Suppose we have any two bases of a finite abelian group G, with each basis element having order a power of a prime. Then the two bases have the same number of elements and the orders of the elements of one basis are, in some arrangement, the same as the orders of the elements of the other basis.*

PROOF. To prove part (ii) of this theorem, suppose that one basis of G is given by 8.28. If G has order n, given by 8.25, the order of every element of G is a divisor of n and hence the only possible primes a power of which can occur as the order of any (basis) element are $p_1, p_2, \ldots, p_k$. Suppose that in a second basis of G, the elements $b_1, \ldots, b_t$ are those whose orders are a power of p_1. Then the elements of G whose elements are a power of p_1 are precisely the elements of the direct sum

8.30 $$(b_1) \oplus \cdots \oplus (b_t).$$

It follows that the direct sum 8.30 is equal to $G(p_1)$. Since from 8.28 we also have

8.31 $$G(p_1) = (a_{11}) \oplus \cdots \oplus (a_{1r_1}),$$

we may apply Theorem 8.21 to the p_1-group $G(p_1)$, and conclude that $t = r_1$, and that the orders of $b_1, \ldots, b_t$ coincide in some arrangement, with the orders of $a_{11}, \ldots, a_{1r_1}$. Thus the number of basis elements having order a power of p_1 is the same in the two bases, as are also the orders of the elements of the two bases. The same argument applies equally well to each prime p_i, and this completes the proof of the theorem.

Let us make the following definition.

8.32 Definition. Let G be a finite abelian group. The orders of the elements of a basis (repetitions being allowed), in which each basis element is required to have order a power of a prime, are called the ***invariants*** (or ***elementary divisors***) of G.

Thus, for example, if we say that G has invariants $3, 2^2, 2, 2$, it means that G is expressible as a direct sum of cyclic groups of these respective orders. Thus for this group G, we have

$$G \simeq C_3 \oplus C_{2^2} \oplus C_2 \oplus C_2,$$

where C_n represents a cyclic group of order n.

The concept of the invariants of an abelian group is important because of the following theorem.

8.33 Theorem. *Two finite abelian groups are isomorphic if and only if they have the same invariants.*

PROOF. One part of this result follows fairly easily from results obtained above. Suppose that $\theta: G \to G'$ is an isomorphism of the finite abelian group G onto the finite abelian group G'. If $\{a_1, a_2, \ldots, a_n\}$ is a basis of G, with each a_i having order a power of a prime, the orders of $a_1, a_2, \ldots, a_n$ are then the invariants of G. Now $\{\theta(a_1), \theta(a_2), \ldots, \theta(a_n)\}$ is a basis of G' (see Exercise 5 below), and under the isomorphism θ, a_i and $\theta(a_i)$ have the some order. Hence G' has the same invariants as G.

Conversely, suppose that G and G' have the same invariants. This means that

$$G \simeq D_1 \oplus D_2 \oplus \cdots \oplus D_n$$

and

$$G' \simeq E_1 \oplus E_2 \oplus \cdots \oplus E_n,$$

where D_i and E_i are cyclic groups of the same order (a power of a prime). Now, by Theorem 6.29, two cyclic groups of the same order are isomorphic. Let $\theta_i : D_i \to E_i$ be an isomorphism of D_i onto E_i. Then it may be verified that the mapping $\theta : G \to G'$ defined by

$$8.34 \qquad \theta(d_1 + d_2 + \cdots + d_n) = \theta_1(d_1) + \theta_2(d_2) + \cdots + \theta_n(d_n),$$

where $d_i \in D_i$ is an isomorphism of G onto G'. (See Exercise 4 below.)

As a simple application of this theorem, let us determine all nonisomorphic abelian groups of order 24. Since the product of the invariants must be 24, we find the following possible systems of invariants: $3, 2^3$; $3, 2^2, 2$; $3, 2, 2, 2$. Thus there are three nonisomorphic abelian groups of order 24. If, as above, we let C_n denote a cyclic group of order n, these three nonisomorphic abelian groups of order 24 are respectively isomorphic to

$$C_3 \oplus C_{2^3}, \quad C_3 \oplus C_{2^2} \oplus C_2, \quad C_3 \oplus C_2 \oplus C_2 \oplus C_2.$$

Let us make explicit a few facts which are implied by the previous results. Suppose G is a cyclic group of order n. Then by Theorem 6.30, every subgroup of G is also cyclic; in particular, $G(p)$ is cyclic for every prime divisor p of the order of G. Since $G(p)$ is cyclic, there is just one invariant of G which is a power of p and, if n is given by 8.25, there are exactly k invariants of G—namely, the numbers $p_i^{e_i}$ for $i = 1, \ldots, k$.

Now we may give a characterization of cyclic groups.

8.35 Theorem. *A finite abelian group G is cyclic if and only if for each prime p dividing its order, G has exactly one subgroup of order p.*

PROOF. If G is cyclic, then there is one subgroup for each divisor of its order and so, in particular, there is just one subgroup of order p for each prime p dividing its order.

Conversely, suppose that G is a finite abelian group with just one subgroup of order p for each prime p dividing its order. Then for each such p, $G(p)$ has only one subgroup of order p. If $G(p)$ has two or more invariants, p^m, p^s, each greater than 1, then $G(p)$ has a subgroup $(a) \oplus (b)$ with (a) of order $p^m > 1$, and (b) of order $p^s > 1$. Each of these has a subgroup of order p which conflicts with the assertion that there is only one subgroup of order p. Hence $G(p)$ has just one invariant. It follows that the invariants of G are exactly the same as the invariants of a cyclic group having the same order as G. By Theorem 8.33, G must be isomorphic to a cyclic group, and hence G is itself cyclic.

We shall use this characterization in the next section.

EXERCISES

1. If an abelian group G has invariants 2^2, 5, 5, verify that there exist elements G of order 20 and none of higher order. Determine the number of elements of order 20.
2. In the notation used in the proof of Theorem 8.21, prove that the set 8.22 is a basis of G_p.
3. In the notation of the same proof, prove that the set 8.23 is a basis of pG.
4. Verify that the mapping θ defined in 8.34 is an isomorphism of G onto G'.
5. Prove: If $\{a_1, a_2, \dots, a_n\}$ is a basis of an abelian group G and if $\theta: G \to G'$ is an isomorphism of G onto G', then $\{\theta a_1, \theta a_2, \dots, \theta a_n\}$ is a basis of G', and G and G' have the same invariants.
6. If $p_1, p_2, \dots, p_k$ are distinct primes, show that any two abelian groups of order $p_1 p_2 \cdots p_k$ are isomorphic (and therefore isomorphic to the cyclic group of this order).
7. Suppose that an abelian group G has order $n = p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k}$, where the p's are distinct primes and each $e_i \geq 1$. If among the invariants of G the highest powers of these primes which occur are $p_1^{t_1}, p_2^{t_2}, \dots, p_k^{t_k}$, prove that there exists an element of G of order $p_1^{t_1} p_2^{t_2} \cdots p_k^{t_k}$ and no element of higher order. [See Exercise 14 at the end of Section 7.4. Observe also that Exercise 1 above involves a verification of a special case of this result.]
8. Verify that there are exactly four nonisomorphic abelian groups of order 100. For each of these groups, determine the maximal order of an element.
9. Show that if a cyclic group G has order p^m (p a prime) and if $t \in \mathbf{Z}$ such that $0 \leq t \leq m$, then G has a subgroup of order p^t.
10. Prove that if p and t are as in the preceding exercise, any abelian group G of order p^m has a subgroup of order p^t. [Hint: Consider the invariants of G, the result of the preceding exercise, and Exercise 2 of the preceding set.]
11. Use Theorem 8.2, Lemma 8.27, and the results of the two preceding exercises to prove the following general result: If an abelian group G has order n and $k \mid n$, then G has a subgroup of order k.
12. If an abelian group G has order n and $k \mid n$, then G has a homomorphic image which is an abelian group of order k.
13. Let G and H be finite abelian groups which have the following property: For each positive integer k, the number of elements in G which have order k equals the

number of elements in H which have order k. Prove that G and H are isomorphic. [Hint: Prove G and H have the same invariants.]

14. The *partition function*, $P(n)$, is defined for positive integers n as the number of ways of writing n as the sum of one or more positive integers with two sums being counted as the same if they differ only in the order in which summands are written. For example, $P(5) = 7$ because

$$\begin{aligned} 5 &= 4 + 1 = 3 + 2 = 3 + 1 + 1 = 2 + 2 + 1 = 2 + 1 + 1 + 1 \\ &= 1 + 1 + 1 + 1 + 1. \end{aligned}$$

Show that the number of nonisomorphic groups of order p^n, for a prime p, is $P(n)$.

8.5 THE MULTIPLICATIVE GROUP OF A FINITE FIELD

Now that we have proved results about finite abelian groups, we are able to complete some of the ideas which were studied in Sections 5.5 and 5.6. For a field F, we denote by $F^\times$ the set of nonzero elements of F. Then $F^\times$ is a multiplicative group.

8.36 Theorem. *Let F be a finite field. Then the multiplicative group $F^\times$ is cyclic.*

PROOF. Suppose that F has order r (necessarily a power of some prime) so that $F^\times$ is a finite abelian group of order $r - 1$. We shall prove there is only one subgroup of $F^\times$ having order q for each prime q dividing $r - 1$. If H is a subgroup of $F^\times$ having order q, then every element in H satisfies the equation $x^q = 1$, 1 being the identity of F. By Theorem 4.46, the polynomial equation $x^q - 1 = 0$ has no more than q solutions in F. Thus there cannot be another subgroup of $F^\times$ having order q, for if there were such a subgroup, its elements would also be roots of this polynomial. Now Theorem 8.35 implies that $F^\times$ is a cyclic group.

This can be used to give some additional information about finite fields and irreducible polynomials.

8.37 Theorem. *Let p be a prime and F a field of order p^n. Then there is an element $\alpha \in F$ such that $F = \mathbf{Z}_p[\alpha]$. Moreover, this α is the root of an irreducible polynomial of degree n in $\mathbf{Z}_p[x]$.*

PROOF. Let α be a generator of the cyclic group $F^\times$. Then every nonzero element of F is a power of α and so the field $\mathbf{Z}_p[\alpha]$ must equal F. The kernel of the evaluation map $\theta_\alpha : \mathbf{Z}_p[x] \to F$ is generated by an irreducible polynomial which has degree, d, say. Then, by the discussion of Galois Fields in Section 5.5, we know that $\mathbf{Z}_p[\alpha]$ has p^d elements. Since $\mathbf{Z}_p[\alpha] = F$ has p^n elements, it follows that $d = n$, and α is the root of an irreducible polynomial of degree n in $\mathbf{Z}_p[x]$.

8.38 Corollary. *For any prime p and any positive integer n, there is an irreducible polynomial of degree n in $\mathbf{Z}_p[x]$.*

PROOF. For any given prime p, and any positive integer n, there is a field with order p^n, by Theorem 5.41; so by Theorem 8.37, there is an irreducible polynomial of degree n.

EXERCISES

1. Let F be a finite field of order p^n, p a prime. For any positive integer r, show that there is an irreducible polynomial of degree r in $F[x]$. [Hint: Consider the field K, of order p^{nr}. Use Theorems 5.44 and 5.50 to argue that it may be assumed that $F \subseteq K$. Now use an evaluation map $\theta_\alpha : F[x] \rightarrow K$ just as in the proof of 8.37.]
2. Let L be any field (not necessarily finite) and let G be a *finite* subgroup of the multiplicative group $L^\times$ of nonzero elements of L. Show that G is cyclic.

Bibliography

Many of the books listed below contain bibliographies that may be used to supplement this limited one. The books listed here should, in large part, be readable by the student of this text. Almost all include topics not included in this book, and collectively they will give the interested student a variety of approaches to the subject.

Abstract Algebra in General

Allenby, R. B. J. T., *Rings, Fields and Groups: An Introduction to Abstract Algebra,* Edward Arnold, London, 1983.

Fraleigh, John B., *A First Course in Abstract Algebra,* 3rd ed., Addison-Wesley, Reading, Mass., 1982.

Herstein, I. N., *Topics in Algebra,* Blaisdell, New York, 1964.

McCoy, Neal H., *Fundamentals of Abstract Algebra,* Allyn and Bacon, Boston, 1972.

Advanced Abstract Algebra

Jacobson, Nathan, *Basic Algebra, I and II,* W. H. Freeman and Co., San Francisco, 1980.

Group Theory

Hall, Marshall, Jr., *The Theory of Groups,* Macmillan, New York, 1961.

Rotman, Joseph, *An Introduction to the Theory of Groups,* 3rd ed., Allyn and Bacon, Boston, 1984.

Ring Theory

Herstein, I. N., *Noncommutative Rings* (Carus Mathematical Monograph No. 15), Mathematical Association of America, Washington, D. C., 1968.

McCoy, Neal H., *The Theory of Rings,* Macmillan. New York, 1964.

Index

A

Abel, 189
Abelian group, 189
Absolute value
  in an ordered integral domain, 56
  of a complex number, 84
Addition
  of complex numbers, 79
  of cosets, 153
  of natural numbers, 23
  of polynomials, 94
Additive group of a ring, 187
Additive inverse, 25
Alternating group, 222
Angle of a complex number, 84
Archimedean property
  of rational numbers, 74
  of real numbers, 77
Associates, 128
Associative law, 24, 25, 187
Associative operation, 21
Associativity of mappings, 194

B

Basis of abelian group, 255
Bijection, 15*n*
Binary operation, 20
Boolean ring, 45

C

Cancellation law
  in groups, 190
  of addition, 36
  of multiplication, 53
Cartesian product, 10
Cauchy's Theorem, 242
Cayley's Theorem, 204
Center, 241
Centralizer, 247
Characteristic of a ring, 163
Closed under an operation, 20
Common divisor, 111
Commutative law
  of addition, 24
  of multiplication, 25
Commutative operation, 21
Commutative ring, 26
Complex numbers, 79–91
Conclusion, 4
Congruent, 142, 152
Conjugate
  of a complex number, 81

conjugate (*cont'd.*)
  of a group element, 236
  subgroup, 236
Constant term, 97
Contrapositive, 3
Converse, 3
Coset, 153, 213
Cycle, 217
Cyclic group, 207

D

Definitions, 2
Degree of a polynomial, 96
De Moivre's Theorem, 86
Derivative, 126
Direct product, 192
Direct sum
  external, 252
  internal, 252
  of abelian groups, 251
  of rings, 31
Discriminant, 178
Disjoint cycles, 218
Distributive laws, 25
Division Algorithm
  for Gaussian integers, 105
  for integers, 104
  for polynomials, 99
Divisor, 98, 107, 111
Divisor of zero, 53
Double coset, 217
Double root, 125

E

Eisenstein's Theorem, 176, 178
Elementary divisors, 263
Empty set, 7
Epimorphism, 48*n*
Equality
  of mappings, 13
  of polynomials, 93
  of sets, 7
Equivalence
  class, 19
  relation, 18
Equivalent implications, 3
Euclidean
  function, 102
  domain, 102
Euclidean Algorithm
  for integers, 114
  for polynomials, 109
Euler Phi Function, 160–161
Euler's Theorem, 161, 216
Evaluation map, 120
Even permutation, 222
Exponent, 43, 60
Extension of a field, 67
External direct sum, 252

F

Factor
  group, 225
  of an element, 98, 111
  ring, 154
  Theorem, 122
Fermat's Little Theorem, 161, 216
Field, 65
  extension, 67
  finite, 163–174
  of complex numbers, 79–82
  of quotients, 67, 70, 72
  of rational numbers, 72, 73–75
  of real numbers, 75–78
  splitting, 140, 147
15-Puzzle, 228
Finite order, 208
First Isomorphism Theorem, 155
Fixed point, 240
Fundamental Theorem
  of algebra, 182–184
  of arithmetic, 131
  of finite abelian groups, 262
  on group homomorphisms, 226

G

Gauss, 180
Gaussian integers, 104, 134
Galois field, 167
Generator
  of a cyclic group, 207
  of a principal ideal, 150, 151
Greatest common divisor (g.c.d.)
  in Euclidean domains, 112
  in polynomial rings, 107
Group
  abelian, 186
  alternating, 222
  cyclic, 207
  dihedral, 248
  octic, 200
  of permutations, 198
  of rigid motions, 199–200
  symmetric, 196

H

Homomorphic image, 48, 201
Homomorphism
  of groups, 201
  of rings, 47, 48
Hypothesis, 4

I

Ideal, 149
Idempotent, 162
Identity
  for an operation, 21
  of a group, 187
  mapping, 15
Image, 12
Implication, 3
Independent elements, 254
Indeterminant, 92
Index, of a subgroup, 215
Indirect proof, 4
Induction Principle, 59
Injection, 15*n*
Integers, characterization of, 62
Integral domain, 53
Internal direct sum, 252
Intersection of sets, 8
Invariant subgroup, 224
Invariants of an abelian group, 263
Inverse
  additive, 25, 36, 37
  implication, 3
  in a group, 187
  multiplicative, 38
  of a mapping, 15
Irrational number, 76
Irreducible
  element of an integral domain, 126, 133
  polynomial, 115
Isomorphic image, 48, 202
Isomorphism
  of groups, 202
  of rings, 48

K

Kernel of homomorphism, 154, 204

L

Lagrange's Theorem, 215
Leading coefficient, 97
Least upper bound, 76
Left coset, 213*n*, 224
Left ideal, 149
Left unity, 45
Length of a cycle, 217
Lexicographic order, 64
Linear combination, 109

M

Mapping, 12
Mathematical induction, 58–61

Modulo, 142, 152, 157
Monic polynomial, 107
Monomorphism, 48*n*
Multiple, 44, 98
Multiplication
  of complex numbers, 79
  of cosets, 153, 224
  of natural numbers, 24
  of polynomials, 94
Multiplicative group of a field, 188, 266–267
Multiplicative inverse, 38, 39
Multiplicity of a root, 125

N

Noncommutative ring, 26
Normal subgroup, 224
Normalizer, 244, 247
*n*th root, 78, 88

O

Octic group, 200
Odd permutation, 222
One-one mapping, 14
One-to-one correspondence, 16
Onto mapping, 14
Order
  of a group, 208
  of an element, 208
  of a field, 165
Ordered
  field, 73
  integral domain, 54
  pair, 10
  ring, 57

P

Partition, 20
Partition function, 266
Permutation, 195
  even, 222
  group, 196, 198
  odd, 222
p-group, 256
Pointwise
  addition, 27
  multiplication, 27
Polynomial, 92
Positive element, 54
Postulates, 1
Power set, 11
Prime
  element of an integral domain, 126, 127, 133
  integer, 130
  polynomial, 117
  subfield, 164
Primitive *n*th root, 91
Principal ideal, 150
Principal *n*th root, 78
Product of mappings, 193
Proper
  subgroup, 191
  subset, 7

Q

Quotient field, 72

R

Rational number, 26
Real number, 76
Real polynomial function, 183
Recursive definition, 43
Reflexive property, 18
Relation, 17
Relatively prime, 110, 114
Remainder, 101, 102
Right coset, 213*n*, 224
Right ideal, 149
Right unity, 45

Ring, 25
  Boolean, 45
  of integers modulo *n*, 157–162
  of matrices, 29, 30
  of polynomials, 92
  of subsets, 31
Root, 120

S

Second Isomorphism Theorem, 157
Set, 6–11
Simple move, 229
Standard form of an integer, 131
Subfield, 67
Subgroup, 191
  normal, 224
  Sylow, 246
Subring, 40
Subset, 7
Surjection, 15*n*
Sylow's Theorem, 244
Symmetric group, 196, 217–223
Symmetric property, 18

T

Transitive property, 18
Transposition, 219
Trichotomy law, 54
Trigonometric form, 84

U

Undefined terms, 1
Union of sets, 8
Unique Factorization Theorem
  for Euclidean domains, 128
  for polynomial rings, 117
Unit, 112
Unity, 26, 39, 45, 46
Upper bound, 75, 76

V

Venn diagram, 8

W

Well-defined, 70
Well-ordered, 57
Wilson's Theorem, 162

Z

Zero of a ring, 25, 35

OTHER IMPORTANT PUBLICATIONS

Burton, David M	Elementary Number Theory 2/e
Coposon, E T	Metric Spaces
Davis, H F et al	Introduction to Vector Analysis 5/e
Dugundji, James	Topology
Hardy, G H	A Course of Pure Mathematics 10/e
Maddox, I J	Elements of Functional Analysis 2/e
Whitaker, E T	A Course of Modern Analysis 4/e

Rs 150

ISBN 81-85392-71-4